

Taking lessons from the new philanthropists

Philanthropies are invigorating US biomedical science by liberating talented researchers from the bureaucracy that surrounds traditional peer review. The NIH should consider distributing a small proportion of its funds in a similar way.

The truism that diversity is strength is borne out by the current health of the biomedical research enterprise in the United States. The past few years have seen a surge in philanthropic funding for biomedicine. Entrepreneurs who made their fortunes in high-tech industry have started giving their spare millions to leading researchers (see page 140), and established players such as the Howard Hughes Medical Institute (HHMI) have continued to grow.

With the budget of the National Institutes of Health (NIH) also on the rise (see page 134), there is much for bench researchers to celebrate. But the new philanthropists have done more than simply increase the resources available for biomedical research. They have also provided an alternative to the bureaucracy associated with traditional grant applications. Rather than asking researchers to fill in forms running to 25 pages or more, and subjecting these to careful, but cumbersome, peer review, some philanthropists have adopted an aggressive strategy of picking potential winners. For the researchers singled out, the experience is liberating. The treadmill of chasing federal research dollars is replaced by a refreshing focus on producing results.

It would be foolish to suggest that the NIH should abandon its traditional 'study sections', through which peer-review panels consider the merits of competing proposals. This system may be conservative and slow, but it is the most accountable way of ensuring quality control that we know. In a period of long-term budget growth, however, the NIH should consider whether some of its riches might be distributed

using similar methods to those adopted by the new philanthropies.

There is otherwise a danger that the NIH will find itself becoming the second port of call for the most talented biomedical researchers. The HHMI supports only around 350 investigators, but these individuals account for a disproportionate number of the most highly cited papers. And if one of Silicon Valley's multimillionaires is prepared to give you hundreds of thousands of dollars to pursue a novel line of research, why bother waiting around for NIH funding?

There is also the long-term health of US biomedicine to consider. The NIH could argue that established foundations and the new philanthropies are already ensuring a diversity of modes of funding. But philanthropists can take away, as well as give. High-tech companies are already feeling the chill winds of an economic slowdown. And as their founders begin to feel the pinch, there can be no guarantee that they will continue pouring money into biomedical research labs.

In rising to this challenge, the NIH does not even have to invent a model of federal research funding. It could simply follow the example of the Defense Advanced Research Projects Agency (DARPA). This body hires temporary programme managers from academia, and gives them almost complete freedom to fund innovative, high-risk projects. Despite its lack of obvious accountability, Congress tolerates this model because DARPA contributes to national security. Do the potential gains for human health not justify spending a few hundred million dollars of the NIH's budget in a similar way? ■

A portal for physics

Nature is introducing a new online gateway into a flourishing and diverting discipline.

Physicists may not have been smiled on by the advisers who drew up US President George W. Bush's first budget request, but their discipline is going through exciting times, nevertheless.

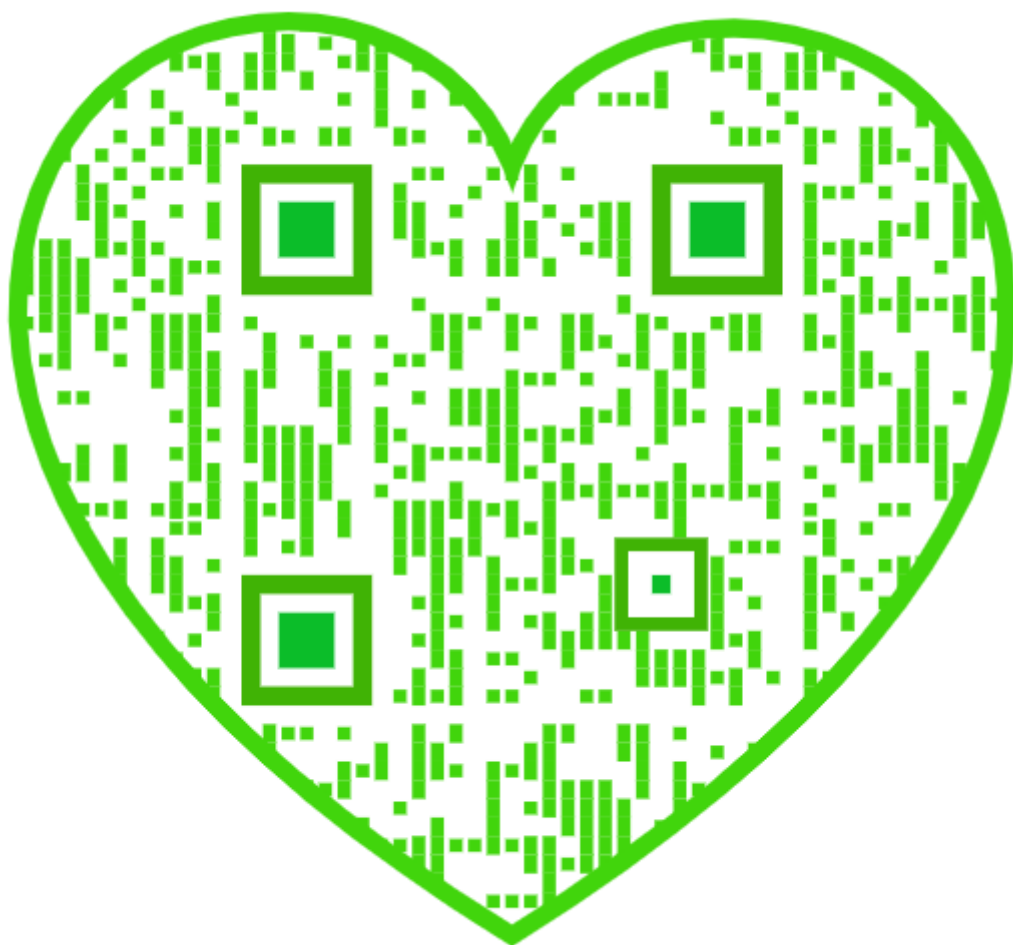
Last week, for example, *Nature* published the report of the discovery of a surprising new superconductor, magnesium diboride (see *Nature* **410**, 63–64; 2001). This is thought to work by a conventional mechanism, but has an unprecedentedly high transition temperature for this type of superconductor, and seems not to suffer from the limitations that have afflicted the high- T_c copper oxide materials (see page 186). Next week's meeting of the American Physical Society in Seattle is already attracting a buzz of excitement not dissimilar to the tumult that followed the discovery of the copper oxides. Thanks to the unpredictable character of condensed matter, the rule in this area of physics seems to be that delightful surprises can be confidently expected — leaving new technologies flowing in their wake.

Other areas of physics are also flourishing. Device physics is feeding the revolution in communications and computing, with spintronics, single-electron devices and photonics all making rapid headway. Computational physics is helping to bridge the gap between micro-

and macroscopic behaviours of simple and complex systems. At the fundamental level, Bose–Einstein condensates are opening up new territory for exploring quantum mechanics. And that is just to focus on 'small' physics — the big-science disciplines of high-energy physics and astrophysics are similarly thriving.

Nature has enjoyed its own resurgence in physics content in recent years, following a dearth dating back to the Second World War. Now we are going a step further. On 12 March we launch the *Nature* Physics Portal (<http://physics.nature.com>), intended to ease access not only to original research published in *Nature*, but also to that appearing elsewhere.

There are other features too — including some high-level fun. We hope the 'problem page', which offers real-world problems for physicists (and anyone else interested) to solve, will prove an amusing diversion. The principle here is one imbued in the training of all physicists: that they should be able to answer almost any question on the way the world works, armed with a few basic principles, the skill to combine them with order-of-magnitude numbers, and the ability to think. Maybe this is what makes physicists so valuable to the society that supports them. Perhaps Bush's advisers should take note. ■





Health benefits
President Bush's first budget favours the life sciences
p134



Desert storm
Grant for research into Gulf syndrome triggers controversy
p135



Daresbury saved
New accelerator and imaging centre to rescue threatened lab
p136



Latest genomes
Sequencers to start work on rat and malaria mosquito
p137

Early signs of a thaw in Bush's attitude to global warming

Matthew Davis, Washington

US cabinet-level officials have started stressing the importance of tackling global warming, in what appears to mark a change of direction for the Bush administration. But some environmental groups remain cautious, saying it is too soon to say where Bush will stand on the issue.

The clearest sign of a policy shift came from Christie Whitman, head of the Environmental Protection Agency, who said last week that the government was serious about reducing carbon dioxide emissions.

In an interview with the television network CNN, Whitman said the "science is good on global warming" and, to the extent that carbon reductions will have an impact on global warming, "that's an important step to take". In other interviews, she mentioned the possibility of supporting legislation that limits carbon emissions from power plants.

This would be a marked shift in US policy. The 1997 Kyoto Protocol resulted in a promise by countries to reduce emissions by an average of 5% from 1990 levels before 2012. But the United States has disagreed, with the

European Union (EU) in particular, as to how these reductions should be made. It has resisted reducing its own greenhouse-gas emissions in favour of such measures as planting trees or funding renewable energy projects in developing countries in return for carbon 'credits' against its own emissions.

The United States has disagreed with many countries on how much to offset carbon-absorbing biomass, such as forest and crops, against carbon emissions. This caused November's summit of 150 environment ministers in The Hague to end in deadlock (see *Nature* 408, 503–504).

Whitman said last week that the United States needed time to review its policies before the next talks on implementing the Kyoto Protocol in Bonn. The talks were due to resume in May, but have been pushed back to 16–27 July, in response to a request from the Bush administration.

The review does not necessarily mean the United States will turn its back on the Kyoto agreement (which Bush opposed during his campaign), Whitman told a G8 meeting of environment ministers last week in Italy.



Rosy outlook: Christie Whitman says the US is serious about cutting greenhouse-gas emissions.

One observer close to administration discussions says that Colin Powell, the secretary of state, has been stressing the importance to US–EU relations of addressing climate change. It was on the agenda of his meeting this week with EU officials including Anna Lindh, Sweden's foreign minister. Sweden holds the EU presidency and has made global warming a priority.

Last month, US treasury secretary Paul O'Neill reportedly made a lengthy presentation to a cabinet meeting about the importance of tackling global climate change. Eileen Claussen, president of the non-profit Pew Center on Global Climate Change, where she worked closely with O'Neill, says he "believes it's a serious problem".

Claussen says that, despite the flurry of interest, the Bush administration still seems far from having fixed a strategy. But she finds the early activity promising, given the inattention she felt that climate change received in the last years of Bill Clinton's presidency.

A spokesperson for the European Commission in Washington agrees it is too early to tell what Bush's tack will be, but the commission hopes he "will share our sense of urgency in dealing with climate change".

'Cheap solution' for climate change

Quirin Schiermeier, Munich

Greenhouse-gas emissions could be stabilized cheaply, but inertia is impeding a change to energy-saving policies and technologies. This is the main conclusion of a new report by the Intergovernmental Panel on Climate Change (IPCC).

The IPCC cost-analysis of climate protection — or mitigation — policies estimates that industrialized countries could achieve the Kyoto Protocol's emission targets at a cost of no more than 2% of gross domestic product; perhaps much less.

"It will now be essential to convince governments that the costs of climate protection are significantly lower than some analysts argue," says Ogunlade Davidson, co-chair of the mitigation working group.

Some 200 lead authors from 120 countries contributed to the mitigation

analysis. They conclude that the long-term economic benefits of climate protection will exceed the costs of using climate-friendly technologies, such as solar and wind power.

In two previous reports, the IPCC had provided strong scientific evidence for a rapid increase in mean global temperatures, and warned that global warming could have devastating consequences (see *Nature* 409, 445 and 971; 2001).

The new report makes few specific recommendations other than "careful consideration of the [environmental and economic] consequences".

Some environmental groups, such as the UK-based Global Commons Institute and the Worldwatch Institute, criticize what they call the vagueness of the analysis, its emphasis on economic factors, and its lack of concrete recommendations.

NIH emerges victorious in Bush's first budget

Irwin Goodwin, Washington

The US National Institutes of Health (NIH) emerged as the sole winner among federal science agencies in President George W. Bush's first budget request to Congress last week. The request calls for an overall 4% increase in public spending to a record \$1.96 trillion.

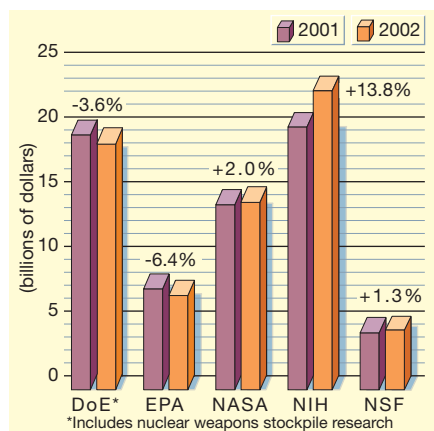
But several agencies, including the National Science Foundation (NSF), the Department of Energy (DoE), NASA and the Environmental Protection Agency (EPA), are left with no increases — and in some cases reductions — in funding.

The NIH is in line for a large funding boost, as expected (see *Nature* **409**, 969; 2001). Its requested 13.8% increase of \$2.8 billion on its current budget of \$20.3 billion leaves it on track to double its 1998 appropriation by 2003.

The NSF can only be envious. In September, Rita Colwell, the NSF's director, requested a 15% increase to the agency's \$4.4 billion



Bare bones of the budget: Bush enlists support for his spending plans at Fernbank Museum in Atlanta.



Disappointing for many: Bush's 2002 budget.

budget. But its ambitions to boost programmes in such areas as nanotechnology, terascale computers and earthquake simulation will be compromised by Bush's budget, which would provide only a 1.3% increase.

The DoE had also asked for a 15% increase on its \$19.7 billion budget. Instead it faces a cut of 3.6%. Bush's 2002 blueprint also calls for increased spending on aid for energy-saving improvements to houses, tax credits for solar and renewable fuels, and some \$275 million extra for stewardship of the nuclear weapons stockpile.

The DoE had hoped to increase grants to university researchers and to fund more users of facilities in high-energy and nuclear physics, materials sciences and fusion research, as well as expanding its genome and bioengineering research at Los Alamos and Brookhaven. The proposed budget reduction puts paid to such plans, and is likely to postpone refurbishment at ageing national labo-

ratories such as Los Alamos and Oak Ridge.

But one senior DoE official insists that its pledges to Europe's Large Hadron Collider, the high-energy accelerators at Fermilab, Brookhaven and Stanford, and the Spallation Neutron Source under construction at Oak Ridge in Tennessee, will not be affected.

NASA, slated for a 2% increase to \$14.5 billion, has been specifically told to cancel its Pluto-Kuiper Express and Solar Probe missions, largely to pay for overruns in the costs of the International Space Station, and to fund a more robust Mars exploration programme. Last week, the agency cancelled the X-33 experimental space plane after investing \$1.3 billion over five years.

But the Bush budget statement is not the last word, says Representative Jim Walsh (Republican, New York), who chairs the House Appropriations Subcommittee. "Ultimately, the Congress will decide what is spent, and that's how it should be."

Indian science benefits from sizeable funding increase

K. S. Jayaraman, New Delhi

Indian science has received a funding boost of 11.7% in the annual budget announced last week. The big winners in the budget, which brings total science spending to 128.4 billion rupees (US\$2.8 billion), are biotechnology, oceanography, and research into solar energy and traditional medicines.

Spending to rebuild the economy of Gujarat, shattered by the recent earthquake, has precluded an across-the-board increase for all ministries.

The Department of Science and Technology will receive an extra 100 million rupees for basic research and 250 million rupees for disaster-mitigation activities following the Gujarat quake.

Industrial research spending will

increase by 5%, and Ragunath Mashelkar, secretary for the Department of Scientific and Industrial Research, is encouraged by the provision of 500 million rupees for "new millennium technology initiatives" and tax relief for pharmaceutical companies entering biotech research.

Atomic energy, space and defence research, which account for the bulk of overall science spending in India, receive rises of 8.9%, 6.4% and 5.8%, respectively. Biotechnology will get an extra 23%, and spending on solar energy will more than triple to 1.74 billion rupees, some of which is earmarked for pilot studies towards a solar electric power plant to be built in the Rajasthan desert.

Funding for oceanographic research rises

by 60%, with money being provided for a joint venture with Russia for the design of unmanned submersibles. These are being developed as part of an Indian programme for recovering manganese nodules from the bottom of the Indian Ocean.

The increase in funding for the Department of Biotechnology is destined for projects in genomics, research into medicinal plants and the creation of an Institute of Bioresource and Sustainable Development.

The budget also provides for the creation of a Traditional Knowledge Digital Library to document traditional remedies — something Mashelkar says may help tell patent officers worldwide about the existence of cures already known in India.

Fracas over \$5 million Gulf syndrome grant

Meredith Wadman

The battle over Gulf War syndrome has broken out again — this time over a US\$5 million grant. The funding has been granted, without peer review, to the laboratory of clinician Robert Haley, whose research on veterans is controversial.

Haley's work suggests that wartime chemical exposure caused brain damage in veterans. US Senator Kay Bailey Hutchison (Republican, Texas) has earmarked \$5 million from the 2001 spending law for the Department of Defense (DoD) specifically for Gulf War illness research at the University of Texas Southwestern Medical Center in Dallas, where Haley works.

"The senator felt Haley was producing some breakthrough research," says a spokeswoman for Hutchison.

But Bernard Rostker, the DoD's special assistant for Gulf War illnesses, disagrees with the decision. "If Senator Hutchison wants special treatment and is willing to explicitly authorize it through legislation, more power to her," Rostker told *Nature*. "But it was not something we [at the DoD] were going to do again after being disappointed the first time."

The \$17 million spent annually by the defence department on Gulf War research is allocated by soliciting proposals on specific



Brain storm: Robert Haley convinced senators to fund him, but some researchers dispute his work.

themes. These are peer-reviewed by panels of independent, non-government scientists. Haley twice failed to win funding in this way. But in 1997, under political pressure to be seen to be taking possible chemical causes of veterans' maladies seriously, the DoD bypassed the process and granted Haley \$3 million. The grant expired last autumn.

Rostker, who supported Haley's original grant, argues that Haley should not "be given special treatment" a second time. He says Haley failed to repeat his original work in a fresh group of veterans who served in the

Gulf and another that did not, as he had agreed to do. Haley disputes this, saying he agreed only to study a new group of Gulf veterans. (He spent \$2.3 million of the grant doing new studies, mainly on his original sample population.)

"We have published the overwhelming majority of the positive literature that has led to information on what's wrong with these [veterans]," says Haley, "and yet we are unable to get funding. You've got politics paralyzing the research process."

Some scientists disagree with this, however. "Dr Haley has basically bypassed the peer-review mechanism," says Philip Landrigan, an epidemiologist at the Mount Sinai School of Medicine in New York, who has written that Haley's studies are methodologically flawed. "That's worrisome. Peer review is a bedrock principle."

Haley published three articles in the *Journal of the American Medical Association* in 1997 (see *Nature* 385, 187; 1997 and 407, 819; 2000). The studies, on 249 members of a Naval Reserve construction battalion that served in the Gulf, identified three neurological "syndromes" and linked them to different self-reported chemical exposures. The most interesting results derived from intensive research on 30 of the veterans.

Critics, including the Institute of Medicine (see *Nature* 407, 121; 2000), have cited the small sample sizes and possible selection bias as flaws in the work. Haley has continued to publish new results using the same smaller group in peer-reviewed journals such as *Archives of Neurology and Toxicology* and *Applied Pharmacology*.

Ross Perot, the Texas billionaire and former presidential candidate, defends Hutchison's actions. "These goofballs in the Pentagon are trying to just sell stress [as the cause of Gulf War illnesses] and not do anything for the men," he says. Perot's private foundation has given Haley \$2.6 million in donations over seven years.

Jury to rule on 'defamatory' paper

Rex Dalton, San Diego

A jury in Arizona is to decide whether the authors of a peer-reviewed paper challenging another scientist's results are guilty of defamation.

The decision to set up a jury trial, made by the state court late last month, is the latest twist in a long-running legal battle involving Ronald Dorn, a geography professor at Arizona State University. Dorn is suing the authors of an article published in *Science* (280, 2132–2139; 1998) in which they questioned his rock-dating methods.

He has settled out of court with one of the paper's co-authors, Wallace Broecker, a prominent Earth scientist at Columbia University, and has opted not to pursue his claims against two in Switzerland.

The remaining co-authors are Warren Beck, Douglas Donahue, Tim Jull and George Burr at the University of Arizona in Tucson, and archaeologist Ekkehart Malotki of Northern Arizona University in Flagstaff.

Dorn's method involves scraping varnish off a rock surface, treating it, and then dating it using radiocarbon analysis. In the *Science* article, the defendants claimed to

have found traces of coal and charcoal in Dorn's samples, which they said could distort the dating results.

At a state court hearing in Phoenix last month, Dorn's attorney argued that the authors of the *Science* paper were involved in "manipulation of the data to make it appear [Dorn] intentionally salted" his samples, which was scientific misconduct.

Paul Carter, the Arizona assistant attorney-general defending the authors, denied the allegations. He noted that some senior science administrators at the University of Arizona judged the article to be within the scope of the researchers' employment. These administrators argued that the manuscript had received extraordinary legal and scientific review before being submitted to *Science*.

Carter sought dismissal of the case on the grounds that the article was a product of research work. But court judge Edward Burke said the dispute over whether the authors "exceeded all of the bounds of acceptable research and commentary" raises a question of fact that can only be decided by a jury. The trial will start in October.



Fresh fields: the new centre 'will offer synergies in treatment, diagnosis and medicine.'

Plan for medical research base secures future of UK lab

David Adam, London

Scientists at the Daresbury Laboratory in northwest England are celebrating the UK government's decision to build a major new accelerator and imaging centre at the lab.

The £150 million (US\$220 million) investment should secure the laboratory's future. This had been in doubt since the government announced plans to build a synchrotron farther south at the Rutherford Appleton Laboratory in Oxfordshire (see *Nature* **404**, 323; 2000) that will cost £550 million to build and run.

Announcing the funding during a visit to Daresbury last week, the trade and industry secretary Stephen Byers said the project would "greatly enhance the northwest's science base, bring developments in the treatment of cancer and improve treatments for other diseases".

"Daresbury has an international reputation for its scientific work," Byers continued, "and is rightly a source of pride both for the region and for the country as a whole."

The new Centre for Accelerator Science, Imaging and Medicine (CASIM) will feature a proton cyclotron — which is used in various research disciplines and in cancer treatment — together with an ion-beam facility for nuclear physics, two free-electron lasers and X-ray imaging equipment. Construction work on the projects, which are still subject to scientific review and feasibility studies, could begin as early as this autumn.

The centre will work in partnership with universities, hospitals and private companies in the region. "It offers clear synergies in treatment, diagnosis and medicine," explains Peter Weightman, a surface scientist at Liverpool University who led the project proposal. Weightman says it will allow

cancer treatments based on X-ray therapy and proton therapy, for example, to be compared for the first time.

Weightman believes that this linking of the science base and the country's health service was key to the proposal's success. The government can dip into both the health-service and core-science funding pots for the money, he says, as well as using cash allocated to the regional development agency. But he insists it should not be seen as regional aid, as the new centre will be of both national and international scientific significance.

Meanwhile, Daresbury researchers are "delighted with the news", says spokesman Tony Buckley. "The staff cheered the secretary of state off the site," he says, adding that workers at the laboratory are determined to make the project a "real scientific success". Some had feared that Daresbury would close when its existing synchrotron radiation source closes in 2007, and many researchers have already left to take up jobs elsewhere.

Byers also announced that a £24 million biopharmaceutical manufacturing facility will be built in the region. It is designed to plug a gap between research and its commercial exploitation by making large numbers of candidate molecules available for early-stage trials.

And, still smarting from last year's synchrotron defeat at the hands of its Oxfordshire rival, the northwest is to get its own 'science council', adding the weight of universities and industry to the lobbying power of its existing regional development agency.

"Not often are scientists lost for words in reaction to a government announcement of funding," Weightman says. "But this morning I think we all were."

Crystal chains invoked as evidence for life on Mars

Tony Reichhardt

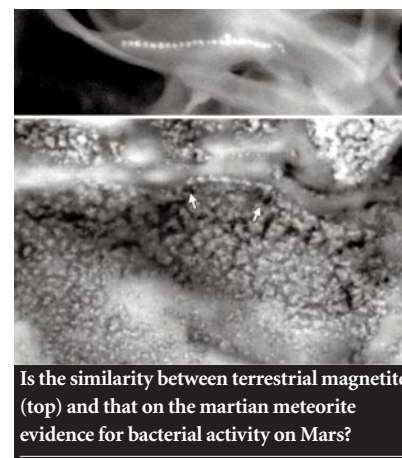
More than four years after researchers at the US space agency NASA reported signs of fossil life in a martian rock, another team has produced what may be the strongest evidence yet to support the controversial finding. But many scientists are still unconvinced by the claims.

Microbiologist Imre Friedmann, an emeritus professor at Florida State University and a specialist on life in extreme environments, says that high-power electron-microscope images of magnetite crystals in the meteorite suggest that they are of biological origin.

Writing in *Proceedings of the National Academy of Sciences* (98, 2176–2181; 2001), Friedmann and his colleagues describe chains of magnetite, which some terrestrial bacteria use to orientate themselves in Earth's magnetic field. They say the chains are too orderly to have been produced without the action of living organisms. The images also reveal possible remnants of membranes around the crystals, which are of uniform size and shape.

Meanwhile, Kathie Thomas-Keptra, a planetary scientist at NASA's Johnson Space Center in Houston, Texas, and co-author of the original paper on fossils in the meteorite, has examined individual magnetite grains. Some 25% of the magnetites in the meteorite show characteristics of crystals produced by living organisms, she and her colleagues write in the same issue of the journal.

But most experts remain sceptical. Allan Treiman, a researcher at the Lunar and Planetary Institute in Houston, will claim in a talk next week at the institute's annual science conference that the grains were created by ordinary decomposition of iron-rich carbonate minerals.



Is the similarity between terrestrial magnetite (top) and that on the martian meteorite evidence for bacterial activity on Mars?

NASA

Postdocs call for better pay and conditions

Corie Lok, Washington

A meeting of postdoctoral researchers and their supporters in Washington last week failed to agree on how to improve pay and conditions. And some questioned whether there were too many postdocs chasing too few tenured positions.

The meeting, organized by the National Academies' Committee on Science, Engineering and Public Policy (COSEPUP), brought together more than 300 researchers, advisers, administrators and representatives of funding agencies and societies.

Their aim was to explore ways of taking forward recommendations made last year in the COSEPUP guide, *Enhancing the Postdoctoral Experience for Scientists and Engineers*.

Postdocs complained that they received inadequate stipends, poor (if any) mentoring, and little support from the administrators or tenured staff at their institutes.

Everyone agreed that change was needed, but not who should take responsibility for it.

"I was disappointed the funding agencies were so refractory to our suggestions by either not answering our questions or saying change was too difficult," said Mary O'Riordan, president of the Postdoctoral Association at the University of California at Berkeley.

Pay was the main concern. A starting salary for a life-sciences postdoc should be in the region of \$35,000 (some \$10,000 more



than at present) said Susan Gerbi, department chair of molecular and cell biology at Brown University in Rhode Island. Marvin Cassman, director of the National Institute of General Medical Sciences and representative of the National Institutes of Health (NIH, the largest supporter of postdoctoral research) agreed that a pay rise was needed.

The researchers called on institutions and funding agencies for help. They asked the NIH and the National Science Foundation (NSF), for example, to include mentoring skills as part of the grant proposal review. But NIH and NSF representatives said they could not enforce this — their role was simply to evaluate research outputs.

Postdocs should find out more before accepting posts, said Jonathan Yewdell, head of the cellular biology section of the National Institute of Allergy and Infectious Diseases. They could assess mentors by talking to people already working in their laboratories.

Institutions should provide supervision and services, said Cassman. At present only a few, such as the University of Pennsylvania School of Medicine, have dedicated offices handling postdoctoral issues. Some are not even sure how many postdocs they have.

"There's an absolute need for a postdoc office," says Tom Sweitzer, co-chair of the Postdoctoral Council at the University of Pennsylvania School of Medicine. "It's the key to solving 90% of the problems."

The researchers themselves questioned whether there were too many of them for the shrinking number of tenure-track positions.

"In the biological sciences, we've got to change our attitudes," says Elizabeth Ann Jones, a postdoc at the Uniformed Services University of Health Sciences in Maryland. "[Tenure] is not likely to happen for all of us."

As many postdocs are actually staff scientists hired to do bench work, one suggestion was to formalize this difference by creating a separate class of 'staff scientists'.

No further meetings have been scheduled yet. Maxine Singer, chair of COSEPUP, said: "We're hoping this will take off on its own". ■

► <http://www.nap.edu/books/0309069963/html/>

Gene sequencers hope to put the bite on mosquitoes

David Adam

The next creatures to have their genomes sequenced will be the laboratory rat and the mosquito chiefly responsible for spreading malaria in sub-Saharan Africa.

The plan to sequence the 260 million base pairs of the mosquito *Anopheles gambiae* was announced this week by a consortium of researchers and sequencing centres.

US company Celera Genomics and Genoscope, the French national sequencing centre, will carry out the initial sequencing using the whole-shotgun technique. They will break the genome into fragments and sequence the DNA of each piece; Celera will reassemble these into a full sequence. Other centres will help to finish and annotate it.

The team expects the project to cost less than \$10 million and be ready, as a rough draft, by the end of the year. The French government will part-fund the effort, and the consortium hopes that US agencies and the European Union will provide the rest.

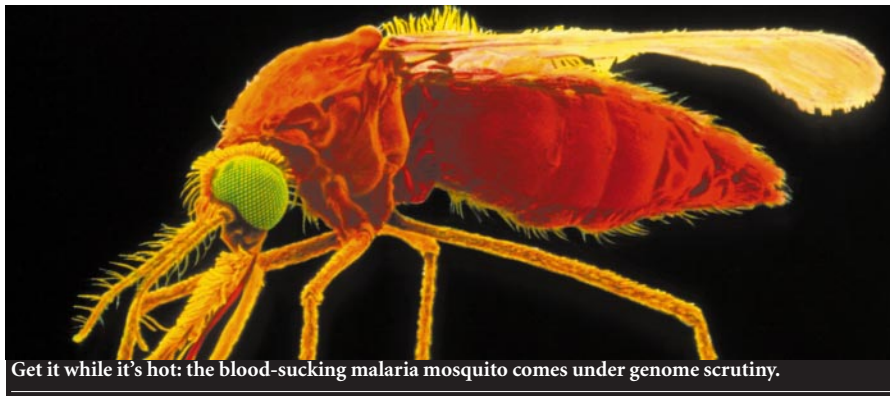
Malaria infects three million people and kills one and a half million each year.

"The sequence will be enormously valuable," says Steven Sinkins, who studies mosquito genetics at the Liverpool School of Tropical Medicine. "It will transform what we can do on a molecular genetics level."

It will help researchers find genes that let mosquitoes carry and transmit the malaria parasite, he predicts, as well as shedding light on insecticide-resistance. Many traits have been loosely traced to genome regions,

but Sinkins says the new work should let researchers focus on individual genes or particular metabolic pathways.

The rat genome — containing an estimated three billion base pairs, similar to the human genome — will be sequenced by Celera Genomics and the Baylor College of Medicine, Houston, sharing a \$58 million grant from the US National Institutes of Health. They aim to finish in two years. ■



Get it while it's hot: the blood-sucking malaria mosquito comes under genome scrutiny.

Arizona's universities reap the rewards of education tax

San Diego The first funds from Arizona's statewide education tax were allocated last week, with \$45 million being earmarked for research projects at the state's three universities.

The Arizona Board of Regents approved research projects and supporting services of \$18.2 million at Arizona State University in Tempe, \$16.3 million at the University of Arizona in Tucson, and \$8.1 million at Northern Arizona University in Flagstaff.

Arizona officials hope the tax will pump as much as \$60 million annually into university research projects involving the life sciences, botany and information technology. California took a similar step last year, using \$300 million in state funds to create three science and technology research centres at the University of California. Other states are considering comparable proposals.

Canada increases support for genome initiative

Montreal Federal funding for Genome Canada, the non-profit-making corporation that aims to make Canada a world leader in

genomics research, has been increased to Can\$300 million (US\$194 million) from the Can\$160 committed in last year's budget to establish the agency.

The bulk of the funds will support research and development in five new regional centres in British Columbia, the prairie provinces, Ontario, Quebec and Atlantic Canada, with 10% set aside to run the national organization. The centres will focus on activities such as functional genomics, genomics sequencing, genotyping, functional proteomics and bioinformatics. Ethical, legal and societal issues related to these fields will also be studied.

Tax breaks aimed at fighting killer diseases

London The British government is to introduce tax incentives aimed at persuading UK-based pharmaceutical companies to boost their research into diseases that affect the world's poor, such as AIDS, tuberculosis and malaria. In an additional move to encourage companies to develop drugs for such diseases, Britain is joining with Italy in backing a new 'global purchase fund' for drugs and vaccines.

Both moves were announced last week by Gordon Brown, the Chancellor of the Exchequer, during an international



Giuliano Amato: backing drug development.

conference in London on child poverty. Italian president Giuliano Amato, addressing the conference by video link, went even further by linking international patent legislation to inequalities in global health spending.

Survey shows public concern over biology

London A survey by the government-funded Human Genetics Commission has revealed that 70% of the UK population has "little or no confidence" that regulations can keep track with and control new research in

biology. The survey also found widespread concern over the use of genetic information by insurance companies and its retention by the police.

As with previous surveys, the poll showed a tension between the perceived benefits of medical science and ideas of what was natural. Of those surveyed, 90% believed that genetic developments should be used to diagnose and provide cures for disease, but roughly a third thought that human genetic research is "tampering with nature, and is unethical".

London college fined over AIDS hazard

London A leading UK university and its safety advisers were fined £20,000 (US\$29,400) last week for exposing the public to a risk of infection from the AIDS virus. The Imperial College of Science in London admitted the breach of safety regulations, but claimed that the risk of exposure was limited to laboratory personnel only.

Imperial College manufactured the virus for research purposes at its laboratory in the Chelsea and Westminster Hospital. Although the court admitted that the chance of the virus escaping from the laboratory was minimal, it found that the possibility of a spillage occurring in an "unsealable" location posed an "unacceptable risk" to the public.

NEAR spacecraft stops transmitting from asteroid

London NASA's much-needed image boost came to an end last week when the NEAR spacecraft finally stopped sending data from its landing site on the Eros asteroid.

The spacecraft surpassed NASA's expectations for the mission when it landed undamaged on Eros and continued to transmit data. NEAR, which stands for Near Earth Asteroid Rendezvous, had already produced spectacular images of Eros during fly-bys in October and November. To the delight of mission scientists, the successful landing enabled NEAR to transmit data from the surface for another 14 days.

The mission goes some way to justifying NASA's policy of 'faster, cheaper, better' missions. The approach took a highly publicized knock last year with the failure of two Mars missions — the Mars Polar Lander and the Mars Climate Orbiter.

Net funding rise for mentor service

Washington The US telecommunications company AT&T has given \$300,000 to MentorNet, a non-profit-making Internet mentoring programme aimed at increasing the number of women in science,

mathematics and engineering (see *Nature* 382, 383; 1996 and *Nature* 391, 11; 1998). The two-year funding is intended to encourage African American and Hispanic women into these fields.

African American and Hispanic women received only about 3–4% of undergraduate degrees in science, mathematics and engineering in 1995. MentorNet was set up to counter this by using e-mail to connect female students with skilled professional advisers. More than 2,000 students from 70 colleges and universities now participate in the four-year-old scheme.

China seeks collaborations with overseas partners

Beijing China is hoping to boost collaborative projects with overseas researchers by setting up a new research centre for information and automation research. The centre will specialize in academic exchanges, technology consultation and the commercialization of research.

The centre is a joint venture between the Chinese Academy of Sciences, Capital Iron and Steel, and the Beijing Association for Science and Technology. Capital Iron and Steel, one of China's leading steel companies, will be given priority in technology transfer projects in return for its support.



Biomedical philanthropy, Silicon Valley style

Entrepreneurs who made their fortunes in high technology are now giving money away to fund biomedical research. These new philanthropists are sending a breath of fresh air through the labs they support, says Trisha Gura.

In the early 1990s, Ben Barres wanted to explore a wild idea: could a long-ignored type of cell play a major role in building the brain's neural circuitry? At the time, most neurobiologists thought these glial cells provided little more than physical support for the brain's neurons.

At his lab at Stanford University in California, Barres painstakingly filled in three consecutive grant applications to the National Institutes of Health (NIH). After each was summarily rejected, Barres was on the point of giving up. But in late 1999, he received word of three reports due to be published in *Nature*. They revealed the identity of a protein — produced by a subset of glial cells — that stunts the regeneration of nerve cells.

The findings were of obvious medical significance, as blocking the protein's function might help treat patients suffering from brain injuries or stroke; they also fitted with Barres's broader thesis that glial cells were doing much more than providing scaffolding within the brain. Barres spent his Christmas holiday co-writing a News and Views article¹ that appeared alongside the three papers²⁻⁴, explaining and celebrating the findings.

Then Barres received a telephone call that changed his working life. The caller was the chief executive of a Californian company that makes microscopes used to check the fabrication of integrated circuits. He had

been fascinated by Barres's article. "He said he'd like to talk to me about the possibility of supporting our work," Barres recalls. "I'd never had a call like that before in my life."

A meeting was swiftly arranged, and Barres is now the recipient of funds generated by a \$3 million endowment. His entrepreneurial benefactor, who wishes to remain anonymous, is an engineer and chemist who holds 17 patents related to nanotechnology.

"I've always been fascinated with the chemical side of the brain," says the donor. He had also seen his mother succumb to dementia.

Barres's story is part of an emerging trend. Having made fortunes in high-tech companies or Internet ventures, many of Silicon Valley's entrepreneurs want to give something back. Some have followed the traditional route of endowing their Alma Mater to set up an institute bearing their



Beneficiaries: Ben Barres (left) and Stephen Strittmatter are funded by the same wealthy individual.



name, or launching a foundation that follows established modes of funding. But others, having grown up in a system where there is no system, are writing their own rules.

Risk takers

These philanthropists have cast aside lengthy grant applications and cumbersome peer-review panels. Instead, funding decisions are placed in the hands of individuals who are told to get to know the scientists working in a field and judge their projects to seek promising and innovative ideas. Sometimes, the benefactors take a hands-on role in deciding where the money should go. The approach has been dubbed 'venture philanthropy', and is the antithesis of the careful but conservative peer review operated by the NIH. By providing seed money to risky projects such as Barres's, the idea is that the winners will eventually amass funding from more traditional sources.

"The money should be viewed as the risk capital in the system," says Queta Bond, president of the Burroughs Wellcome Fund in Research Triangle Park, North Carolina, one of the more traditional biomedical philanthropies. "The new entrepreneurs are good at doing strategic analyses and finding where philanthropy dollars can make a difference. And they do make a difference." Indeed, although the hundreds of millions of dollars being awarded by such philanthropists is easily dwarfed by the \$16.9 billion devoted to extramural research programmes by the NIH for 2001, the speed, flexibility and careful targeting of the new philanthropists' dollars magnifies their actual amount.

In Barres's case, the investment is already bearing fruit. In January, he published a paper⁵ showing that neurons in culture dishes will not form functional synapses, the connections through which they communicate, unless glial cells are also present. Not surpris-

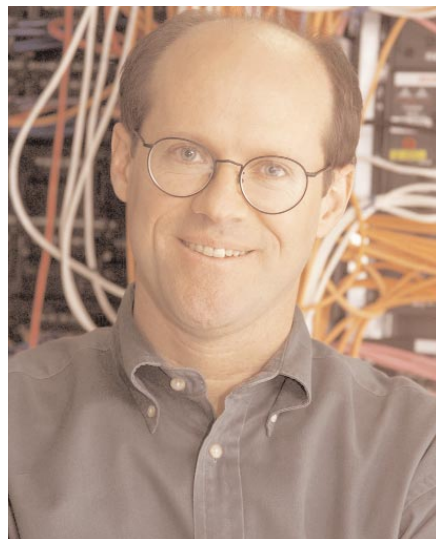
ingly, Barres is wildly enthusiastic about the contribution being made by the new breed of philanthropists. And other beneficiaries are similarly delighted at being freed from the treadmill of writing grant applications. "It lessens the pressures of raising grant dollars from other sources," says Stephen Strittmatter of Yale University in Connecticut, lead author of one of the papers about which Barres wrote his News and Views, and the recipient of funds from the same benefactor.

Big bucks

So who are the new philanthropists, and what are they funding? The most prominent is Microsoft founder Bill Gates, whose Bill and Melinda Gates Foundation, founded in 1994, holds assets worth \$21 billion, and is ploughing hundreds of millions of dollars into targeted health-related programmes — in particular efforts to find vaccines against developing-world killers such as AIDS, malaria and tuberculosis. In January, for instance, the Gates foundation gave \$100 million to the International AIDS Vaccine Initiative (IAVI), providing a major boost to its effort to raise \$550 million to move at least six vaccine candidates into clinical trials by 2007. "I think you have to give Gates enormous credit," says Kenneth Shine, president of the Institute of Medicine in Washington. "By investing as he has in vaccines and infectious disease, he has galvanized others to participate." Indeed, the Internet company Yahoo! announced in January that it was donating \$5 million to IAVI.

But many of the high-tech philanthropists are more interested in supporting research into ageing and neuroscience. "They are interested in how we as humans perceive, use and store information," says molecular biologist Phillip Sharp of the Massachusetts Institute of Technology (MIT), who has been selected to direct the McGovern Institute for Brain Research — for which MIT is receiving \$350 million over the next 20 years from computer publishing mogul Patrick McGovern and his wife Lore Harp McGovern, a high-tech entrepreneur.

Placing cash figures on the rise in biomedical philanthropy by such entrepreneurs is difficult. The Foundation Center in New York, which monitors overall US philanthropic spending, does not break its figures down according to the business background of the donating individuals. But observers



New breed: Steve Kirsch (left) and Sarah Caddick want scientists to have a 'business plan'.

We don't ask for reams and reams of data. We ask: 'Are you thinking along the same lines as us?'

MARY MERRICK

such as the Institute of Medicine's Shine say that the rise in donations from high-tech entrepreneurs has been eye-catching.

In 1998, the Community Foundation Silicon Valley published a survey that provided a portrait of Silicon Valley's new philanthropists. The community foundation is an umbrella body that distributes money on behalf of donors who do not wish to set up their own foundations, but nevertheless want to dictate where the money should go. Its survey showed that the typical donor fell in the 35–44 years age bracket, was newly wealthy, and had an affinity for science.

Peter deCourcy Hero, president of the Community Foundation Silicon Valley, provided a tongue-in-cheek description of these philanthropists in a discussion paper for the Indiana University Center on Philanthropy's 13th Annual Symposium last August: "These are the 40-year-old engineers whose mid-Western school-teacher parents saw Sputnik go up ... [They] said: 'Johnny you become an engineer and you will never starve.' Well, now young Johnny is worth about \$20 million and neither Mom nor Dad left any instructions as to what to do with such a mind-boggling sum."

Gifts for the gifted

Steve Kirsch is one of the new breed. The founder of four successful start-ups — Infoseek, Mouse Systems, Frame Technology and most recently, Propel Software — Kirsch was asked in the early 1990s to serve on a fundraising committee by Leonard Ely, an old-school philanthropist in Silicon Valley. Kirsch was initially perplexed by the request, asking Ely why people donated to such committees. Ely recalls his response: "Steve, believe it or not, some people like giving their money away." He took Kirsch under his wing and taught him about philanthropy.

To begin with, Kirsch deposited his funds with the Community Foundation Silicon Valley, his donations totalling more than \$12 million by 1998. Kirsch's donor-advised fund began supporting individual researchers with interests consistent with his desire to cure or prevent diseases.

But Kirsch, a quintessential Silicon Valley entrepreneur, wanted to do something bigger. "Steve would like to cure all diseases, if he could," says the Kirsch Foundation's Sarah Caddick, who formerly administered the cancer research grant programme of the Damon Runyon–Walter Winchell Foundation in New York. His chance came with a windfall of cash from the sale of a portion of Infoseek to Disney in 1999. The deal gave Kirsch and his wife Michele \$50 million to plough into a new San José-based foundation. By creating the Kirsch Foundation, Kirsch freed himself from the restrictions imposed on donor-advised funds. Foundations, for instance, can engage in political lobbying. "Steve comes from the point of view that you have to use the most

effective moral and legal means to accomplish your goals," says Kathleen Gwynn, the foundation's president.

For Kirsch, 'effective' also meant eschewing the bureaucracy associated with traditional modes of biomedical research funding. Instead of asking for grant applications and putting their review in the hands of committees, Kirsch hired Caddick as director of medical and scientific programmes. Caddick's background in neuroscience fitted with one of Kirsch's priorities, and she is now in charge of figuring out who and what to fund.

Thought experiments

Caddick uses her scientific connections to seek out individuals with promising projects. In addition to reading the literature, she calls people she knows and asks for introductions to people she does not. In short, she networks. "I talk to leaders in the field — Nobel prizewinners, presidents of major institutions, deans and directors of departments," says Caddick. "Often these people don't need money, but they can be most objective about who does."

Instead of a detailed grant proposal, Caddick asks for a 'concept' — an outline of the proposed research in less than two pages. "Assuming that you are a top-notch scientist, I know that you can write a grant proposal," she says. "We don't ask for reams and reams of data. We ask: 'Are you thinking along the same lines that we are thinking of?' If you are, then we can craft a business plan together."

With such talk of business plans, it is clear that the Kirsch Foundation operates more like a Silicon Valley start-up than a traditional foundation. What it wants is creativity with a sound scientific basis: risky, innovative ideas that are likely to draw further funding down the road. Having picked potential winners, the foundation carefully scrutinizes ongoing projects to check that they are delivering the goods. "A researcher may not be able to accomplish in a year what they set out to do," says Caddick. "We just ask that you make some headway in getting there or be able to explain why you cannot." But those that fail this test will find that the plug is pulled. "If you are a scientist not focused on curing diseases, then we are not the right funder for you," says Gwynn.

But for researchers with the appropriate

mindset, the Kirsch Foundation can represent, as Ronald DePinho of Harvard Medical School and the Dana-Farber Cancer Institute puts it, "a match made in heaven". DePinho is interested in the role of telomeres — the 'caps' on the ends of chromosomes — in cellular ageing. Telomeres grow shorter with successive cell divisions, and DePinho wondered if this might explain why diseases involving constant turnover of liver cells, such as hepatitis or alcoholism, often lead to the eventual failure of the organ.

DePinho had gained NIH funding to create mice lacking telomerase, the enzyme that rebuilds telomeres, and last year he published a paper showing that these mice struggle to repair their livers if the organs are damaged, but can be treated by replacing the enzyme⁶. As he prepared this paper for publication, he wanted to push the research forward and realized that getting further NIH funding would entail inevitable delays. So he went to Harvard's development office, leafed through a book of funders, and came up with the Kirsch Foundation. Harvard backed his application for an investigator's award that would pay out \$150,000 annually to DePinho and an additional \$30,000 to the university, for indirect costs. The foundation liked the idea, and started funding DePinho's lab. "They allowed me to seize the moment," says DePinho.

Kirsch anticipates that his foundation will be influential. "I hope that others will adopt the techniques we use to fund medical research," he says. "As we achieve success, this will be much more likely."

The new philanthropies are also cutting through a deceit at the heart of the traditional funding system — the fact that researchers often make their grant applications look convincing by applying for money to conduct research that they have, in fact, already done. They then use the funding to do the research that forms the basis of the next grant application, and so on. The problem lies in producing a competitive grant proposal if you genuinely want to move into a new area. "You can find any number of investigators who will tell you that if you haven't already done the first two years of work, you are in trouble," says Richard Sprott, executive director of the Ellison Medical Foundation in Bethesda, Maryland.

The Ellison foundation, launched in 1998

These are the 40-year-olds whose parents saw Sputnik go up. They said: 'Johnny you become an engineer and you will never starve.' Well, now Johnny is worth about \$20 million and neither Mom nor Dad left any instructions as to what to do with such a mind-boggling sum.



Silicon money: Larry Ellison's (left) foundation is spending \$45 million a year on biomedical research paid for from the fortune he made through his computer company Oracle, based in Santa Clara (above).

by Larry Ellison, chief executive of the computer giant Oracle, shares the Kirsch Foundation's ethos. Its main focus is ageing research — although Ellison recently decided to expand into tropical diseases. Sprott, formerly head of biology programmes at the National Institute of Aging (NIA) in Bethesda, was hired to work out how best to distribute the foundation's budget to complement the NIA's federal dollars. With the addition of its programme on tropical diseases, the foundation is now spending \$45 million a year.

"Our strategy is to pick cutting-edge people and turn them loose," says Sprott. To help find candidates, the foundation has assembled an impressive board, chaired by Nobel laureate Joshua Lederberg of Rockefeller University in New York, and including such luminaries as neuroscientist Eric Kandel of Columbia University, also in New York, who last year shared the medicine Nobel. "There are complaints that people of this calibre don't sit on NIH study sections any more," Sprott says. The reason, he suspects, is that NIH panels "get bogged down in trivia".

Again, paperwork is kept to a minimum. Lederberg's board reviews applications running to no more than four pages, supported by recommendations from colleagues and institutions. Sprott contrasts the Ellison foundation's procedures with the 25-page applications submitted to the NIH.

Wendy Baldwin, deputy director of extramural research at the NIH, defends the agency's peer-review procedures as necessary for quality control — and most scientists would agree that adopting the new philanthropists' techniques across the board would be a recipe for chaos. But even Baldwin admits that the NIH system moves slowly and shies away from more risky proposals. "We are never able to fund all the projects that

ought to get money," she says. Sprott agrees that the NIH and the new foundations are playing different, complementary roles: "We don't need to compete with the NIH," he says.

But for established biomedical foundations, the new philanthropies have brought a little competition. And some are now looking at the overall scene to work out whether they need to refocus their activities. David Seemungal, senior policy adviser at Britain's Wellcome Trust, notes that the new philanthropies tend to operate by "parachute funding" — making large investments in specific research areas with little or no forewarning.

Federal bypass

Recognizing the need for better coordination, philanthropies new and old have recently started getting together to coordinate their activities. In a trend-setting initiative, representatives from the Howard Hughes Medical Institute, the American Cancer Society, the Pew Charitable Trust and the Burroughs Wellcome Fund have over the past 18 months organized two meetings for philanthropists, including the new players from Silicon Valley, to share their experiences. One topic of discussion has been how philanthropies can support research that is denied federal funding as a result of political sensitivities or legal restrictions. Private foundations could, for instance, become an important lifeline for scientists wanting to work with human embryonic stem cells, if the administration of George W. Bush, as expected, denies them NIH funding.

But researchers hoping to gain from the largesse of the new philanthropies have one big worry. Silicon Valley's entrepreneurs owe their immense wealth to the economic boom of the 1990s. With high-tech stocks already having faltered, and with talk of a recession

in the air, there is no guarantee that Kirsch, Ellison and their kind will carry on giving quite so generously. Indeed, neither Kirsch's nor Ellison's foundations have the security of an endowment — both are funded directly with cash from their founders' pockets. And Lori Arthur, who heads Stanford's development office, says that she is planning for a decline in income for the coming year.

But for now, those who have already benefited from Silicon Valley philanthropy urge researchers with an innovative proposal to get straight down to their university's development office, and ask for advice in selecting the donor best suited to fund their particular project. The result, say scientists such as Barres and DePinho, can be liberation from the continual and energy-sapping chase for federal research dollars. "We are buying back the researchers' time," says Caddick. ■

Trisha Gura is a freelance writer in Cleveland, Ohio.

1. Goldberg, J. L. & Barres, B. A. *Nature* **403**, 369–370 (2000).
2. Prinjha, R. *et al.* *Nature* **403**, 383–384 (2000).
3. Chen, M. S. *et al.* *Nature* **403**, 434–439 (2000).
4. GrandPré, T., Nakamura, E., Vartanian, T. & Strittmatter, S. M. *Nature* **403**, 439–444 (2000).
5. Ullian, E. M., Sapperstein, S. K., Christopherson, K. S. & Barres, B. A. *Science* **291**, 657–661 (2001).
6. Rudolph, K. L., Chang, S., Millard, M., Schrieber-Agus, N. & DePinho, R. A. *Science* **287**, 1253–1258 (2000).

Web links

Foundation Center

► <http://www.fdncenter.org>

Bill and Melinda Gates Foundation

► <http://www.gatesfoundation.org>

Community Foundation Silicon Valley

► <http://www.siliconvalleygives.org>

Survey of Silicon Valley philanthropists

► http://www.cfsv.org/giving_back_svway.html

Kirsch Foundation

► <http://www.kirschfoundation.org>

Ellison Medical Foundation

► <http://www.ellisonfoundation.org>

A question of intent: when is a 'schematic' illustration a fraud?

Sir—Scientific images are at the centre of recent allegations of 'bad science' (for example, the 'Indonesian coelacanth' story¹). How can we identify fraudulent alterations in scientific illustrations?

We suggest that the key is to identify changes, then examine them in the light of the scientist's intent. Take, for example, the old dispute between biologists Ernst Haeckel and Wilhelm His about embryos and evolution. Haeckel and others said that different animals pass through, or 'recapitulate', similar embryonic stages². Indeed, fish and human embryos do look similar because they share primitive features — 'symplesiomorphies' in modern terms. Professor His³ disagreed, saying that embryos show distinctive hallmarks (synapomorphies) of their species group.

The evidence on both sides included drawings of embryos^{2,3}. Haeckel's young embryos look similar, whereas His's look different. Things turned nasty when His and others accused Haeckel of doctoring pictures⁴. The defence, even today, is that Haeckel's figures are only schematic: it is acceptable for schematics to show alterations that help to explain the data.

We have found evidence of sleight of hand, surprisingly, on both sides. His's deer embryo has cloven hooves, but a standard work on deer embryology⁵ shows no such feature. This disparity arouses our suspicion. However, we have not identified His's source, so we cannot be sure that he changed anything. His embryos are at more advanced stages than Haeckel's, though, so are not valid counter-evidence.

We can make a persuasive case with Haeckel because we have identified some of his sources. When we compare his drawing of a young echidna embryo with the original⁶, we find that he removed the limbs (see Fig. 1). This cut was selective, applying only to the young stage. It was also systematic because he did it to other species in the picture. Its intent is to make the young embryos look more alike than they do in real life.

Haeckel's other intent is to support recapitulation, as revealed in his text for the young embryos: "There is still no trace of the limbs or 'extremities' in this stage of development ... [this] proves that the older vertebrates had no feet" (ref. 2, p. 371). The altered drawings support theories which the originals did not. Therefore, these are not legitimate schematic figures.

Haeckel and His both published technical works of great importance to biology. Their dubious embryo pictures

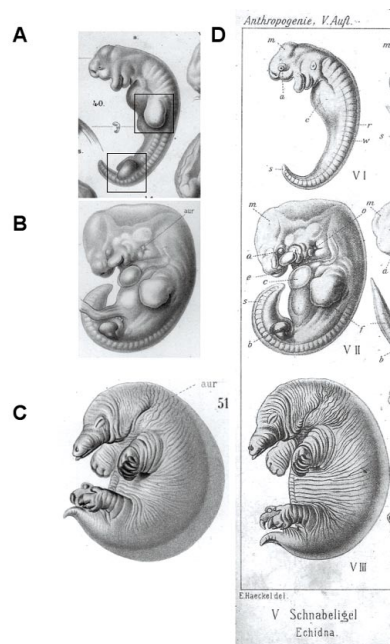


Figure 1 Deleting the synapomorphies from young embryos: Haeckel's echidna embryo. A–C, Tafel X, figs 40s, 43s and Tafel XI fig. 51s, respectively, from the original drawings⁶. D, Haeckel's copy in the *Anthropogenie*, 5th edition (ref. 2 Taf XI, figs VI–VIII). The limb buds of the young embryo (boxed) were deleted by Haeckel. Significantly, the other stages were copied accurately.

appeared in non-technical, polemical books. Ironically, there was some truth on both sides: in fact, embryos show a mixture of symplesiomorphies and synapomorphies⁷.

Michael K. Richardson*, **Gerhard Keuck†**
 *Institute of Evolutionary and Ecological Sciences,
 Leiden University, Kaiserstraat 63, Postbus 9516,
 2300 RA, Leiden, The Netherlands
 †R&D Documentation, Aventis CropScience
 GmbH, Industriepark Höchst, K607, 65926
 Frankfurt am Main, Germany

1. *Nature* **406**, 225 (2000).
2. Haeckel, E. *Anthropogenie oder Entwicklungsgeschichte des Menschen. Keimes- und Stammesgeschichte* (Engelmann, Leipzig, 1903).
3. His, W. *Unsere Körperform und das physiologische Problem ihrer Entstehung* (Vogel, Leipzig, 1874).
4. Gould, S. J. *Natural History* March, 42–49 (2000).
5. Sakurai, T. in *Normentafeln zur Entwicklungsgeschichte der Wirbelthiere* (ed. Keibel, E.) (Fischer, Jena, 1906).
6. Semon, R. *Zoologische Forschungsreisen in Australien und dem malayischen Archipel* (Fischer, Jena, 1894).
7. Richardson, M. K. et al. *Anat. Embryol.* **196**, 91–106 (1997).

Conservation should be a high priority in Singapore

Sir—The Singapore Genomic Project was created in June 2000 with funding of S\$62 million (US\$35 million) over five years¹. Singapore is certainly becoming a hotspot for life sciences — basically to develop drugs, medical and food products and agro-biotechnology. But it is ignoring serious

environmental conservation issues, both at home and in neighbouring countries.

When the state of Singapore was founded in 1819, it had a population of 150 people. The main island (544 km², to which another 30 km² was later added by land reclamation) was almost entirely covered by rainforest. Today, more than half the island is urbanized to accommodate 3 million inhabitants. Less than 100 ha of rainforest and 500 ha of mangrove forest survive, in a degraded state². About 594 of a total of 2,277 native vascular plant flora have become extinct³.

The lack of natural resources means the city-state depends on drinking water from Malaysia and Indonesia, but it makes little financial commitment to conserve natural resources in these countries. Ecology and biodiversity conservation are not popular in Singapore's universities, as the government has been heavily promoting genetics and biomedical-related life sciences and research during recent years. We are worried that the emphasis on the Genomic Project may reduce the number of future students of ecology and conservation in Singapore.

The economy improved in 1999 with surplus of US\$1.96 billion registered⁴. But environmental spending was restricted to deep-tunnel sewerage systems, wastewater treatment and management of nature reserves. International environmental conservation activities were limited to Singapore's participation in meetings.

Being an economic giant in the Asia-Pacific region, Singapore has the responsibility to care for its own shrinking natural environment. It must reorient its strategies and make a financial commitment both to conserve its own environment and to help conserve natural resources in the neighbouring mega-biodiversity countries such as Malaysia, Indonesia and Vietnam⁵.

The government has recently set up Wildlife Reserves Singapore, bringing together three of the world's most progressive animal parks: the Jurong Bird Park, Singapore Zoological Gardens and the Night Safari. It has also established a conservation fund to support wildlife preservation projects locally and globally, owing to the desperate need to promote biodiversity conservation in the region.

It is not too late for other governmental and non-governmental agencies to join this effort to save the rapidly diminishing natural environment in South-east Asia.

Govindasamy Agoramoorthy, **Minna J. Hsu**
 Department of Biological Sciences, National Sun Yat-sen University, Kaohsiung, Taiwan

1. *Nature* **407**, 935 (2000).
2. Corlett, R. J. *Biogeograph.* **19**, 411–420 (1992).
3. Turner, I. M. et al. *Conserv. Biol.* **8**, 705–712 (1994).
4. Government of the Republic of Singapore, Fiscal Development, Annual Report 1999 (Monetary Authority of Singapore, 1999).
5. Mittermeier, R. A., Myers, N. & Mittermeier, C. G. *Hotspots: Earth's Biologically Richest and Most Endangered Terrestrial Ecoregions* (Univ. Chicago Press, 2000).

Falling to Earth in a quantum way

Strings, loops and other tools aiding the quest for unity in the physical world.

Three Roads to Quantum Gravity

by Lee Smolin

Weidenfeld & Nicolson: 2001. 176 pp. £16.99

David Lindley

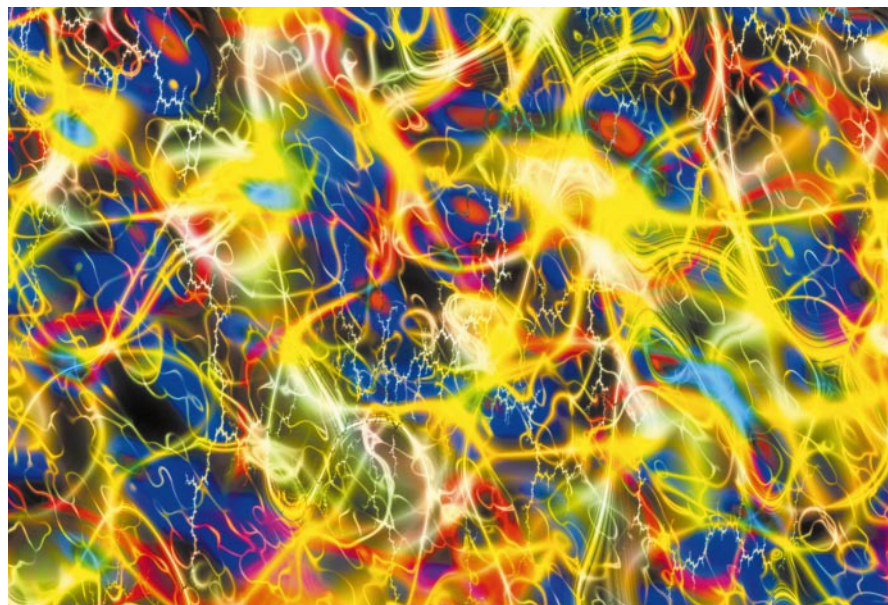
Back in the 1930s, rumour got about that Werner Heisenberg had put together the broad outlines of a theory that would unify gravity and electromagnetism, and that he would publish his work once he had sorted out the details. The ever sardonic Wolfgang Pauli sent a postcard to a friend on which he crudely drew a picture-frame enclosing a blank space, and wrote: "This is to show I can paint like Titian. Only details are missing."

Proclaiming himself an unabashed optimist, Lee Smolin declares that physicists today are within 10 years, 15 at the most, of completing the modern version of Heisenberg's task — the formulation of a theory of quantum gravity. This would tie together the intricacies of quantum mechanics with the decidedly classical world of gravitation and account for all of the known particles and interactions of the physical world. To a cynical fellow such as myself, those are fighting words. But Smolin's eloquent and thoughtful book may persuade many readers that his optimism is not misplaced.

Over recent decades, proclamations of an imminent 'theory of everything' have come from champions first of supergravity, then of string theory. The latter now threatens to evolve into a still unknown entity called, mysteriously, M-theory.

This line of development, which forms one of the three roads of the book's title, flows from the attempt to expand quantum mechanics to include gravitational interactions. As Smolin explains, though, such theories have a fundamental flaw: they rest on the essentially newtonian idea of space-time as an invariable background. They therefore fail to incorporate the link between energy and geometry that is the essence of general relativity.

A less well-known strategy has been to start with relativity and turn it into a quantum theory in which space-time itself would build up from finite, elemental structures. This goal has now been attained in the guise of something called loop quantum gravity, which allows exact treatment of a genuinely quantized geometry. Smolin has spent much of his career on this second road. But he is honest enough to admit that loop quantum gravity also has a flaw: in his words, it's boring. It's a theory of space-time, but nothing else — no nuclear forces, no electrodynamics.



Strung together? Superstrings, represented graphically here, are one attempt at a 'theory of everything'.

There has been a certain antagonism between travellers on these two roads. Each group believes that the other is not attacking the real problem. But, as Smolin cogently explains, it's more constructive to say that each side's accomplishments address the other's failings. The central theme of this book is that a final understanding of quantum gravity must incorporate lessons from both camps, in the form of a theory to which loops and strings emerge as limited approximations from different perspectives.

And that eventual theory, Smolin also argues, will contain ideas from explorers on the third road. These pioneers have come up with profound insights by pondering fundamental questions about space, time and causality. Roger Penrose, for example, devised many years ago a novel geometrical structure called a spin network, which turns out to be closely related to the space-time picture more laboriously generated by loop quantum gravity.

Smolin masterfully explains not only what these various ideas are, but where they come from, why they are important and how they relate to each other. Unlike most writers on these topics, he makes astute use of historical precedent. I particularly liked his argument that black-hole thermodynamics implies that space-time can be quantized in the same way that the statistical definition of entropy more than a century ago pointed directly to the atomization of matter. From time to time the explanations become a little dense and murky, but then Smolin shows a rare ability to bring the reader back on board by cleverly depicting

the essential physics questions that drive theorists' sophisticated explorations.

What of Smolin's professed optimism? His account manages the considerable trick of offering a compelling vision of what a theory of quantum gravity might look like, while plainly admitting that putting all the pieces together remains a formidable task. As Pauli indicated, an obvious concern is that the details *are* the theory, not merely finishing touches that can be worked out at will.

There is also the disturbing question of whether any final theory will be comprehensible to ordinary mortals. Smolin wades into some deep mathematical waters, and once or twice has to confess his difficulty in getting the idea across. There is something called topos theory, for example, which (I think) is a kind of observer-dependent mathematical logic, and which Smolin admits was tough sledding even for him. And despite my best efforts, I couldn't wholly grasp Smolin's portrayal of quantum space-time as an interrelationship of processes and information, rather than the familiar arena for particles and actions.

The implicit hope seems to be that, once all the disparate bricks and planks from the three roads have been hammered together, the scaffolding taken down and the debris carted away, the finished theory will display a readily apparent beauty. I certainly hope so. It would be too bad if theoretical physicists succeed in devising a quantum theory of gravity that they can't explain to anyone else.

But these are reservations about the search for quantum gravity, not about

book reviews

Smolin's excellent portrayal of the quest. In its expansiveness of vision, his book stands far above any other account that I know of for a general audience. Readers may find themselves struggling on occasion, but one can skate over some of the trickier pages and still emerge with a clear picture of the intended goal. That picture is, in one or two places, quite breathtaking in its elegance and power.

Whether Smolin's optimism turns out to be justified, only time (or whatever time becomes, in quantum gravity) can tell. Not all physicists, he admits, share his sunny outlook, and others will disagree entirely with his view of where this research is going. Nevertheless, Smolin has made a wonderfully persuasive case that the quest for quantum gravity is not just alive, but positively humming. ■

David Lindley, author of *Boltzmann's Atom: The Great Debate That Launched a Revolution in Physics* (Free Press, 2001), is a writer currently living in Arlington, Virginia, USA.

Follow the Building Block Road

The Wizard of Quarks: A Fantasy of Particle Physics

by Robert Gilmore

Springer: 2000. 202 pp. £14.95, \$24

Christine Sutton

When Dorothy sets off on a subway ride in the Big City with Uncle Henry and Aunt Em, she little expects to find herself in a land of witches, with a scarecrow, a tin man and a lion for company. Even less does she expect to learn about wavefunctions, amplitudes, nuclear binding energy and gauge theory.

If the setting sounds familiar, that is no mistake, for Robert Gilmore has chosen L. Frank Baum's children's story, *The Wizard of Oz*, as an allegory of our present understanding of the elementary particles of matter and the forces that act on them. The result is a strange (pun intended) *mélange* of corny humour and rather laboured analogy, interlaced with some excellent non-mathematical explanations of a broad range of sub-atomic physics.

Like her earlier namesake, Gilmore's Dorothy makes a journey through an imaginary land populated by remarkable characters. But, rather than learning to appreciate home life in Kansas like the original Dorothy, she finds out about many of the major discoveries of modern physics. From an encounter with the Witch of Mass ("You may call me G"), via the Wizard of Quarks, the travellers follow the Building Block Road until they arrive at the Planck Energy.

Their route mirrors the twentieth century's journey from atoms to quarks, during

which particle physicists have discovered that matter is more peculiar than they could ever have imagined. Tiny quarks lie imprisoned within protons and neutrons, to be freed only in the company of additional quarks (or anti-quarks), in the guise of new particles. The quarks (and the seemingly unrelated leptons) interact through forces that are understood in terms of further particles, the gauge bosons, which act like balls in a game of quantum catch. So does it help to have the fabulous reality of particles and forces explained in terms of a storybook fantasy?

There is little to fault in Gilmore's non-mathematical descriptions of difficult physical concepts, and he succeeds in covering a great deal of ground, from atomic spectra to the Standard Model. But the descriptions — whether voiced by Dorothy's erudite companions or set out in separate explanatory paragraphs — are not for someone with no previous knowledge of the basic ideas of atomic physics and quantum theory. A story that has introduced the four fundamental interactions and Planck's constant by page 13 is not for the faint-hearted.

Similarly, the attempts at humour will often be lost on a reader who does not already know the physics, and the simple punning begins to wear thin — for example, the half-man, half-horse that the travellers meet in the Kingdom of CERN is the Visitor Information Centaur.

So, if not for the complete beginner, will the allegory work for someone who already knows a little modern physics? Even then, there is a danger that the effort to recall Judy Garland's celluloid progress down the Yellow Brick Road to the Emerald City will distract from the careful explanations of concepts in theoretical physics.

This is the third book by Gilmore that



Can Dorothy's celluloid progress work as an allegory for the discoveries of modern physics?

seeks to explain modern physics through characters from familiar stories; readers previously joined *Alice in Quantumland* and were treated to *Scrooge's Cryptic Carol*. So it must be an allegorical style that works — but for whom? Or does this reader lack a sense of humour? Bah, humbug! ■

Christine Sutton is in the Nuclear and Astrophysics Laboratory, University of Oxford, Keble Road, Oxford OX1 3RH, UK.

Subtended by evolution

Who Wrote the Book of Life? A History of the Genetic Code

by Lily E. Kay

Stanford University Press: 2000. 441 pp. \$60, £37.50 (hbk), \$24.95, £15.95 (pbk)

Horace Freeland Judson

The title is the give-away. The idea of the book of life, or more generally the book of nature, goes far back, of course. Charles Darwin began the *Origin of Species* by quoting Francis Bacon from the early seventeenth century: "Let no man out of a weak conceit of sobriety, or an ill-applied moderation, think or maintain, that a man can search too far or be too well studied in the book of God's word, or in the book of God's works." The classic notion was that God wrote both books, and therefore that what we call scientific research is not heretical. But that's not the meaning of Lily Kay's title. The authors of Kay's book of life are the scientists whose competition put together the genetic code.

Who Wrote the Book of Life? attempts two things, which cannot be disentangled. First, and centrally, it presents research into the origins of molecular biology in the 1940s and 1950s, culminating in a reconstruction of the intense competition to break the genetic code — that is, to identify the exact list of the 64 three-base sequences (codons) in DNA that, by way of RNA intermediates, specify the 20 amino acids essential to proteins. It is best to be explicit here, for the definition of the genetic code is one of Kay's concerns.

Some of the research is Kay's own, from interviews and her scrutiny of scientists' notebooks. But Kay has heroically synthesized her own findings with a vast mass of published literature. And she sets this competition in the context of the press and public excitement it generated. The resulting narrative is in some ways new and in some passages thrilling, especially the account of that intense competition. Characters involved include Marshall Nirenberg and Heinrich Matthaei, who in 1961 co-discovered the first two codons, Severo Ochoa who discovered many more, and a host of more minor players, with Francis Crick setting the



Francis Crick (left) and Claude Shannon: pioneers of two very different kinds of information transfer.

theoretical framework and scolding them for their sloppiness.

Kay has gone rather overboard in her admiration of Nirenberg. His desperate ambition and *faux-naïve* enthusiasm displayed in his notebooks and later interviews evidently resonated with her. She undervalues the achievements of Har Gobind Khorana, who shared the Nobel prize with Nirenberg in 1968.

But she sets her best material in two larger, theoretical contexts. The simpler of these is the rise, in the 1940s and 1950s, of information theory — the mathematical study of how information can be transmitted as streams of sequences of symbols. But did information theory really play an important part in the rise of molecular biology and in particular in the elucidation of the genetic code during those same years?

Kay produces impressive evidence to show that molecular biologists knew about information theory and were familiar with its terminology. Yet, as she indeed explains, this theory had its roots in the wartime mathematical analyses of cipher-breaking by Claude Shannon and others, and in Shannon's attempts after the war to apply the analysis to the transfer of signals, such as voice transmission by telephone and the transmission of television images.

But this sort of information is in bulk, treated statistically. It entails the ideas of entropy — the degradation of order — and of redundancy, necessary to restore order. A few physicists in the early 1950s tried unsuccessfully to apply such ideas to biological processes.

The meaning of information in molecular biology is totally different: no bulk messages, but rather individual sequences and their related control elements. Information in that sense was defined by Crick with elegant parsimony in his celebrated talk delivered to the

Society for Experimental Biology in September 1957, where he formulated the sequence hypothesis and the Central Dogma: "The Central Dogma ... states that once 'information' has passed into protein *it cannot get out again*. ... Information here means the *precise* determination of sequence, either of the bases in the nucleic acid or of amino acid residues in the protein."

So, yes, this is information, but not information theory *sensu stricto*. Furthermore, for biologists both at that time and today, the term is a simple metaphor — useful shorthand, but no divining-rod. Despite Kay's elaborations, the differences between the two meanings of the word information remain distinct and opposed. In some passages she acknowledges this.

Then, that other context. At this point, the book becomes difficult to discuss. This is partly because Kay's treatment is extremely complex and the terminology and systems of analysis she uses are dense and specialized, even hermetic. But the difficulty arises also because a coterie of historians of science whom she tried with only limited success to impress during life will now flock to her defence. For Lily Kay died last December, of cancer. Kay was luminously optimistic — lean, hair scraped back tight, face scrubbed, eyes shining — and was a woman of great and unaffected courage at all levels. The people she wanted to please can be characterized loosely as the post-modernist, social-constructionist historians of science.

And so Kay attempts to place her story of the genetic code in that fashionable context. We get references to Lucretius and St Thomas Aquinas, to the I Ching and Chomsky. Suddenly in the final chapter we get lengthy discussion of the linguist Roman Jakobson. We get aporias, instantiations, aphoristic energy, autopoiesis, and more of the same.

And we get repeated attempts to distinguish meanings — notably, four different meanings of 'the genetic code', when in fact these are no more than four aspects of one idea.

Yet Kay also delivers cautions and concessions. Towards the end of the book there is the wonderful observation by François Jacob, the most rigorously intellectual of all the founders of modern genetics: "Language studies the messages transmitted from an emitter to a recipient. Now there is nothing of the kind in biology: no emitter, no recipient. The famous message of heredity transmitted from one generation to the other, no one has ever written it; it is constituted by itself, slowly, painfully traversing the vicissitudes of reproductions subtended by evolution." Here, then, is the answer to the question in Kay's title. ■

Horace Freeland Judson is at the Center for History of Recent Science, George Washington University, Washington DC 20052, USA.

A rare diversion

The Best American Science Writing 2000

edited by James Gleick

Ecco Press: 2000. 258 pp. \$14 (pbk)

Rogene M. Eichler West

It is difficult enough for scientists to find the time to keep up with the technical publications in their field, much less to squeeze in the consumption of a book that can't be cited in a grant proposal. But *The Best American Science Writing 2000* may be worth making an exception for.

For those aspiring to science writing (essays by researchers and journalists are fairly evenly represented), this collection contains a variety of styles that might be categorized and dissected to reveal their underlying principles of form and function.

A number of pieces typify familiar genres: specifically, biographies, the recasting of one's research into layman's terms, and news reporting. Yet other pieces challenge conventional classification. One example is "When doctors make mistakes" by the physician and writer Atul Gawande. Gawande escorts the reader into the operating theatre of an emergency room during a crisis, into the amphitheatre where the entire medical staff meets to scrutinize the events of the past week, and through the ongoing process of uncovering and correcting systemic causes of error. Although his story is, at one level, a tale of process and chance, it is by providing the reader with glimpses into the culture of the medical profession, its rites and rituals, that his story acquires the power to convince.

In "Einstein's clocks: The place of time",

science historian Peter Galison recounts the story of Europe's attempts to solve a practical turn-of-the-century problem, namely, the normalization and synchronization of clocks. By alternately reviewing the progress of this effort with the temporally concurrent progression of young Albert Einstein's career as a Swiss patent-office clerk, Galison implicitly reminds the reader of the degree to which science and scientists are influenced by and develop within the context of those seemingly unrelated events unfolding around them.

In yet another historical revisit, "A division of worms", Stephen Jay Gould corrects the record of a man whose legacy has long been associated with a discredited theory of evolution. The reader learns that Jean-Baptiste Lamarck's contributions to biology (the name that he, incidentally, gave to the field) have profoundly influenced the structure of the modern-day evolutionary tree. But more importantly, Gould's sleuthing reveals a man who, late in his career, had the courage to abandon the guiding theory of his life's work to accommodate the underlying order suggested by a new set of observations.

A few selections disappoint. The offering from the satirical newspaper *The Onion* is "Revolutionary new insoles combine five forms of pseudoscience". Although a clever and well-written parody of science news reporting, its composition did not require any particular deftness for science writing *per se*, which brings into question its inclusion in this collection. In the case of writer Susan McCarthy's "Must dog eat dog?", an awkward mix of facts and facetiousness result in an unbalanced pop-science style that seems out of place alongside smooth-flowing selections by such notable science popularizers as Oliver Sacks and Timothy Ferris.

An additional criticism of this collection lies with what appears to be a bias in the process of selecting the works: four of the 19 selections were originally published by *The New York Times*. This would not be an issue were it not for the fact that editor James Gleick was formerly an employee of that newspaper, and that no other newspaper, aside from *The Onion*, has work represented. With all due respect to the extraordinary talent at *The New York Times*, reporters from, for example, the *Chicago Tribune* have this past decade had a more successful record of Pulitzer prizes for explanatory (science) journalism.

These criticisms may warrant the dismissal of the superlative "Best" from the book's title. But the volume itself should not be dismissed. It is rare to be offered such a diverse collection of science writing, even more, one that can be enjoyed by laymen, scientists and writers alike.

Rogene M. Eichler West is at *Science Renaissance*, PO Box 21946, Seattle, Washington 98111, USA.

Science in culture

Localized lumps

Phrenology as serious science.

Martin Kemp

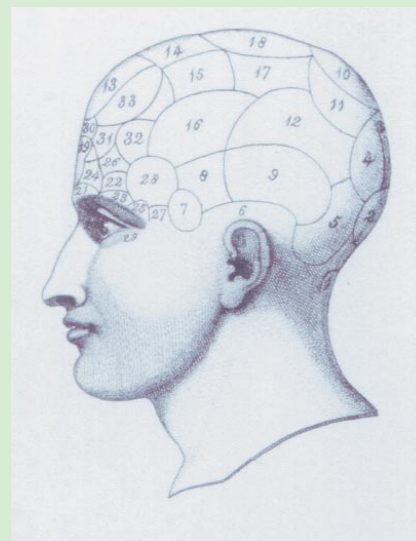
Localization is all the rage in brain science. Driven by vivid and often colourful imaging technologies, modern neurology is progressively disclosing the 'hot' areas of local activity associated with particular mental processes.

The quest to identify which centres in the brain are responsible for particular functions dates back at least to classical antiquity, most notably to Aristotle's highly influential *De anima*. In the Middle Ages, Aristotle's speculations were codified in terms of faculty psychology, in which imagination, common sense, voluntary and involuntary action, intellect and memory were assigned locations in progressively deeper ventricles of the brain.

The most substantial later initiative to place faculty psychology on an empirical basis was forged during the Enlightenment by the science of phrenology. Now commonly stigmatized as a pseudo-science, akin to physiognomics and chiromancy and worth at most a few humorous side-swipes in modern texts on neuropsychology, phrenology was founded on logical extrapolations from the most advanced research into brain structure.

Phrenology was created by Franz Joseph Gall, an Austrian physician, and promoted in a series of heavy-duty publications written with his follower, Johann Gaspar Spurzheim. The most famous of these was *Anatomie et physiologie du système nerveux général*, published in four substantial volumes with an "Atlas" in 1810–1819. Opening with a massively comprehensive review of the science and philosophy of brain and mind, Gall and Spurzheim subsequently provided extensive data on the relationship between cranial shape and mental faculties in the human and animal kingdoms. Studies of developing brains and crania encouraged them to believe that the cranium was specifically shaped to accommodate the variously configured brains in different species, and even in individuals of the same species. Gall declared, not unreasonably, that "the form and size of the brain regulate the form and size of the skull".

Drawing on the "general law" that "throughout all nature, the properties of bodies act with an energy proportional to their size", Gall searched for any prominences in the globe of the cranium that might betray highly developed features housed within. The problem was that he had no direct access to the brain activity, and his only recourse was to correlate data on people possessing special attributes with unusually prominent 'bumps'. Spurzheim recalled that "if the head of any individual presented any protuberance, which was evidently the result of cerebral development, Gall endeavoured to be acquainted with the talents or dominant



Mapping the mind: a phrenological head with localization as devised by Gall and Spurzheim.

character of the person". The method was, therefore, impeccably empirical in the eighteenth-century sense, and the anatomical premises were far from daft.

The misfortune of Gall and Spurzheim was not just to be wrong in their detailed explanations — which is the historical fate of much science in the long term — but to be taken up in the public domain in a form that laid their ideas open to ridicule. The reading of the 'bumps' of the brain to diagnose people's characters became the speciality of ill-informed opportunists. The founders were appalled that their serious science had been adopted as an "art of prognostication".

"We consider only the faculties man is endowed with, the organic parts, by means of which these faculties are manifested, and the general indications which they present. The object of this new psychological system is to examine the structures, the functions and the external indications of the nervous system in general, and of the brain in particular. Thus does this science especially contribute to the knowledge of human nature."

The only fundamental difference between this declaration and principles of modern neurology is their emphasis on "external indications" — but when it came to processes in the living brain, external signs were all that phrenologists could reliably access.

From this historical perspective, one might wonder whether the large claims that are being made for modern neurology are not in some danger of suffering from the kind of skewed public reception that so distorted the foundational ideas of phrenology.

Martin Kemp is in the Department of the History of Art, University of Oxford, 59 George Street, Oxford OX1 2BE, UK.

When words fail

Scientists have to struggle with words that don't fit reality.

Frank Wilczek

Language is a social creation. It encodes the common experience of many people, past and present, and has been sculpted mainly to communicate our everyday needs. Ordinary language is most certainly not a product of the critical investigation of concepts. Yet scientists learn, think and communicate in it during much of their lives. Ordinary language is therefore an unavoidable scientific tool — rich and powerful, but also quite imperfect.

One scientific imperfection of language, perhaps the most obvious, is its incompleteness. For example, there are no common words for several of the most central concepts of quantum theory, such as the linearity of state-space and the use of tensor products to describe composite systems. To be sure, we've developed some applicable jargon — 'superposition' and 'entanglement', respectively, are the words we use — but the words are unusual ones, not likely to convey much to outsiders, and their literal meaning is misleading to boot.

Although it creates cultural barriers and contributes to the balkanization of knowledge, such enrichment and slight abuse of language is not conceptually problematic. Much more insidious, and more fundamentally interesting, is the opposite case: when ordinary language is *too* complete. When something has a name, and is commonly employed in discourse, it is seductive to assume that it refers to a coherent concept, and an element of reality. But it need not. And the more pervasive the word, the more difficult it can be to evade its spell.

Few words are more pervasive than 'now'. According to his own account, the greatest difficulty Einstein encountered in reaching the special theory of relativity was the necessity to break free from the idea that there is an objective, universal 'now': "[A]ll attempts to clarify this paradox satisfactorily were condemned to failure as long as the axiom of the absolute character of times, *viz.*, of simultaneity, unrecognizedly was anchored in the unconscious. Clearly to recognize this axiom and its arbitrary character really implies already the solution of the problem."¹

Einstein's original 1905 paper begins with a lengthy discussion, practically free of equations, of the physical operations involved in synchronizing clocks at distant points. He then shows that these same operations, implemented by a moving system of observers, leads to differing determinations

of which events occur "at the same time".

As relativity undermines 'now', quantum theory undermines 'here'. Heisenberg had Einstein's analysis specifically in mind when, in the opening of his seminal paper on the new quantum mechanics in 1925, he advocated the formulation of physical laws using observable quantities only. But while classical theory has a naïve conception of a particle's position, described by a single coordinate (a triple of numbers, for three-dimensional space), quantum theory requires this to be replaced by a much more abstract quantity. One aspect of the situation is that if you don't measure the position, you must not assume that it has a definite value. Many successful calculations of physical processes using quantum mechanics are based on performing a precise form of averaging over many different positions where a particle 'might be found'. These calculations would be ruined if you assumed that the particle was always at some definite place. You can choose to measure its position, but performing such a measurement involves disturbing the particle. It changes both the question and the answer.

Einstein himself was never reconciled to the loss of 'here'. In his greatest achievement, the general theory of relativity, Einstein relied heavily on the primitive notions of events in space-time and (proper) distance between nearby events. These notions rely on unambiguous association of times and places — 'nows' and 'heres' — to individual objects of reality (though not, of course, on the existence of a universal 'now'). Understandably impressed by the success of his theory, Einstein was loath to sacrifice its premises. He resisted modern quantum theory, and did not participate in its sweeping success in elucidating problem after great problem.

Ironically, the sacrifice he feared has not (yet) proved necessary. On the contrary, in the modern Theory of Matter, we retain 'nows' and 'heres' for the fundamental objects of reality. These primitives are no less important in the formulation of the subatomic laws of quantum theory than in general relativity. The new feature is that the fundamental objects of reality are one step removed from the directly observed: they are quantum fields, rather than physical events.

It is possible to avoid ordinary language and its snares. Within specific domains of mathematics, this is accomplished by constructing exact definitions and axioms. Purity of language is also forced on us when we interact with modern digital computers,



Time trial: Einstein found it hard to break free of the idea of 'now'.

CORBIS

since they do not tolerate ambiguity.

But the purity of artificial languages comes at a great cost in scope, suppleness and flexibility. Perhaps computers will become truly intelligent when they learn to be tolerant of ordinary, sloppy language — and then to use it themselves! In any case, for us humans the practical and wise course will be to continue to use ordinary language, even for abstract scientific investigations, but to be very suspicious of it. Along these lines, Heisenberg's considered formulation, put forward in *The Physical Principles of Quantum Theory* in 1930, was: "[I]t is found advisable to introduce a great wealth of concepts into a physical theory, without attempting to justify them rigorously, and then to allow experiment to decide at what points a revision is necessary."

Looking to the future, after 'now' and 'here', what basic intuition will next require reformation? As the nature of mind comes into scientific focus, might it be 'I'? Perhaps the following remarks of Hermann Weyl, stimulated by deep reflection on the aspects of modern physics discussed here and stated in his *Philosophy of Mathematics and Natural Science* (1949), point in that direction: "The objective world simply is, it does not happen. Only to the gaze of my consciousness, crawling upward along the life line of my body, does a section of this world come to life as a fleeting image in space which continuously changes in time." ■

Frank Wilczek is in the Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

1. Einstein, A. "Autobiographical notes" in *Albert Einstein, Philosopher-Scientist* (ed. Schilpp, P.) (Library of Living Philosophers, 1949).

Room for doubt

Roger A. Pielke Jr

Who really killed John F. Kennedy? How many whales inhabit the oceans? How will climate change? Does God exist? Strive as we might, answers to such questions cannot be known with certainty. Uncertainty means that more than one outcome is consistent with our expectations. Expectations are a result of judgement, sometimes based on technical mistakes and interpretive errors, and shaped by values and interests. Because uncertainty is a characteristic of every important decision, it is no surprise that society looks to science and technology to help clarify our expectations in ways that lead to desired outcomes.

Decision-making is forward-looking; consequently, decision-makers in governments and other organizations have traditionally looked to science and technology to quantify and if possible reduce uncertainties about the future. In many cases, particularly those associated with closed systems — or systems that can be treated as closed — understanding uncertainty is a straightforward technical exercise: probabilities in a card game, error analysis in engineering and manufacturing, or the actuarial science underlying many forms of insurance, for example. But in many other circumstances, systems are not closed, and understanding uncertainty in an open system is considerably more challenging. Many scientists have taken on the challenge of understanding such open systems — global climate, genetic engineering, and so on. And the process of securing the considerable public resources to pursue this challenge often results in their explicit promise to “understand and reduce uncertainties”.

Conventional wisdom holds that uncertainty is best understood or reduced by advancing knowledge, an apparent restatement

of the traditional definition of uncertainty as “incomplete knowledge”. But in reality, advances in knowledge can add significant uncertainty. For example, in 1990 the Intergovernmental Panel on Climate Change (IPCC) projected that a doubling of CO₂ in the atmosphere would result in a 1.5 to 4.5 °C mean global temperature change. In 2001, after tens of billions of dollars of investment in global-change research, the IPCC now concludes that a doubling of CO₂ will result in a 1.5 to 6.0 °C temperature change. Even as the IPCC has become more certain that temperature will increase, the uncertainty associated with its projections has also increased. Why? Researchers have concluded that there are many more scenarios of possible population and energy use than originally assumed, and have learned that the global ocean-atmosphere-biosphere system is much more complex than was once thought. Ignorance is bliss because it is accompanied by a lack of uncertainty.

Science and technology can also make the certain uncertain. Consider, for example, the invention of nuclear power, chlorofluorocarbons and genetically modified organisms (GMOs). In seeking to reduce uncertainties related to power generation, refrigeration and agriculture, each invention created new uncertainties: radiation disasters, ozone depletion and ecosystem response, respectively. From this perspective, a key difference between the critics and the champions of science and technology is that the former focus on new uncertainties (the glass half-empty), the latter on reduced uncertainties (the glass half-full).

The apparently successful policy response to chlorofluorocarbon-induced ozone depletion rests on an effective mix of science and technology and decision-making, with the role of the latter under-appreciated. Without careful attention to this mix — and

Uncertainty

“Consensus science can only provide an illusion of certainty.”

particularly its effects on our expectations — society may take the responsibility for reducing uncertainty out of the hands of science and technology, as has been the case with nuclear power (and seems to be occurring with GMOs), possibly limiting potential benefits from new technologies for ever.

Consider once again global climate change. For many years, policy debate has centred on the degree of certainty that decision-makers ought to attach to competing visions of the future climate. Lost in this doomed enterprise is the point that climate will certainly have an increasingly strong effect on the environment and society, simply because of growing vulnerability related to factors such as population, wealth and use of land. If a goal of climate policy is to reduce the effects of climate on the environment and society, then effective action need not wait until we are more certain about details.

Seen in this light, efforts to reduce uncertainty via ‘consensus science’ — such as scientific assessments — are misplaced. Consensus science can provide only an illusion of certainty. When consensus is substituted for a diversity of perspectives, it may in fact unnecessarily constrain decision-makers’ options. Take for example weather forecasters, who are learning that the value to society of their forecasts is enhanced when decision-makers are provided with predictions in probabilistic rather than categorical fashion and decisions are made in full view of uncertainty.

As a general principle, science and technology will contribute more effectively to society’s needs when decision-makers base their expectations on a full distribution of outcomes, and then make choices in the face of the resulting — perhaps considerable — uncertainty. ■

Roger A. Pielke Jr is at the National Center for Atmospheric Research, PO Box 3000, Boulder, Colorado 80301, USA.

FURTHER READING

Bankes, S. C. Exploratory modeling for policy analysis. *Operations Research* **41**, 435–449 (1993).
Oreskes, N., Shrader-Frechette, K. & Belitz, K. Verification, validation, and confirmation of numerical models in the earth sciences. *Science* **263**, 641–646 (1994).
Pielke, R. A. Jr & Betsill, M. M. Policy for science for policy: ozone depletion and acid rain revisited. *Research Policy* **26**, 157–168 (1997).
Sarewitz, D., Pielke, R. A. Jr & Byerly, R. (eds) *Prediction: Science, Decision-making, and the Future of Nature* (Island Press, Washington DC, 2000).

Degrees of uncertainty: more research has led to wider projections for the effect of CO₂ on temperature.

Weighing the Universe

Marc Davis

How do you measure the mass of the Universe? You can't use a balance to weigh it, but detailed analysis of hundreds of thousands of galaxies provides an alternative answer.

Cosmology has historically been considered a branch of philosophy rather than physics because of the dearth of data. But there has been dramatic progress in the past few years, and it is now entering an era of large-scale studies and precise measurements. The latest dispatch from this new frontier is the interim report by Peacock and collaborators on page 169 of this issue¹. By systematically studying not hundreds or thousands of galaxies, but hundreds of thousands of them, they have produced a comprehensive map of the local Universe, which allows them to measure some of the fundamental parameters that define the structure of the entire cosmos.

For more than 70 years, astronomers have known that the Universe is expanding and that the velocity with which a galaxy appears to move away from us is approximately proportional to its distance. This smooth 'Hubble expansion' is direct evidence that the Universe began at a finite time in the past. By studying the spectrum of light from a particular galaxy we can accurately measure its velocity along the line of sight by the magnitude of the resulting Doppler effect, or redshift as it is known, thereby providing a measure of its distance. During the past 20 years, with the advent of sufficiently sensitive detectors and computer-controlled instruments, detailed mapping of the heavens in three dimensions has become possible. Figure 1 shows a three-dimensional slice of the nearby cosmos². This detailed map shows the galaxy distribution to be highly inhomogeneous, with large voids, long filamentary structures stretching in excess of 100 million light years (10²¹ km), and dense clusters of galaxies where the filaments cross.

Peacock *et al.*¹ now present a new map of nearby galaxies up to three billion light years away (see Fig. 1 on page 170). These are the first precise measurements from the 2dF collaboration — a group of British and Australian astronomers using a specially designed instrument at the Anglo-Australian telescope in Australia to study 250,000 galaxies in detail. The resulting map has similar texture to that seen in Fig. 1 here but covers a much larger volume. There is a dramatic contrast between the highly structured clustering of galaxies revealed by these maps and the uniform Universe inferred from measurements of the radiation in the cosmic microwave background^{3,4}

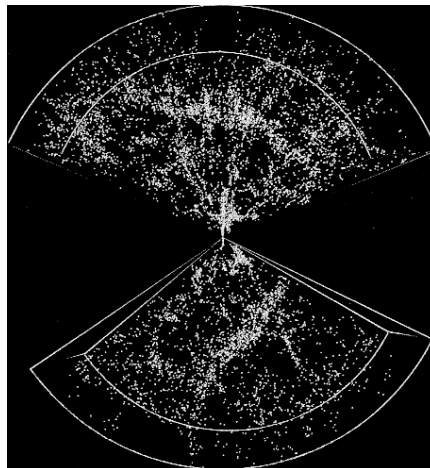


Figure 1 A slice of the local Universe, in which each galaxy is plotted as a point². The Earth is located at the centre of the plot, and the outer boundary is 400 million light years away (corresponding to a redshift of 0.04). The galaxy distribution appears to be highly inhomogeneous, with large voids and long filamentary structures stretching across 100 million light years. Scientists from the 2dF survey¹ have now produced a much larger version of this map (out to 3 billion light years), which allows them to estimate the mass density of the Universe from the pattern of clustering of some 140,000 galaxies.

(CMB). The CMB is a relic from the hot, dense and uniform phase of the initial Big Bang. So how did the striking lumps and bumps that we see as galaxies or clusters of galaxies today develop from such a homogeneous early Universe?

Theorists have long suspected that structures are formed by a process known as 'gravitational instability'. In this picture, the expansion of slightly denser regions in an otherwise smooth, expanding Universe is slowed by gravity more rapidly than the surroundings, increasing the contrast in mass density between high and low density regions. But finding direct evidence in support of this reasonable conjecture has been difficult. The timescale for the growth of galaxy structures (clustering) is the age of the Universe itself, far too long for direct observation. But because the strength of clustering is expected to increase with time, the galaxies must deviate from the smooth Hubble expansion. These deviations away from uniform Hubble flow are known as 'peculiar

velocities', and they come in two forms. The first instance occurs within the compressed centres of dense clusters of galaxies. Here, the general outward push of the Universe has long since been overwhelmed by the local tug of newtonian gravity, and these galaxies have high peculiar velocities driven by the local gravitational potential.

The second type of deviation has been seen by Peacock *et al.*¹. They have convincingly shown that peculiar velocities can also be detected for larger-scale structures that are still expanding. Whereas the peculiar motions within clusters of galaxies are incoherent and random, the clustering in long filamentary chains and other dispersed structures causes opposing sides of the structure to move coherently towards each other, partially cancelling the Hubble expansion. So galaxies on the near side of a collapsing structure are falling away from us, increasing in velocity compared with the rate of Hubble expansion, whereas objects on the far side fall towards us, reducing their velocity. We can detect the Doppler shift of this velocity only along our line of sight, so if we map galaxies using redshift to indicate distance, structures that have not yet collapsed completely will appear to be compressed slightly along our line of sight. Such regions should show up against the background of randomly oriented structures.

Peacock *et al.* report the first convincing detection of this effect in their analysis of the statistical properties of the clustering of galaxies along and perpendicular to the line of sight (see Fig. 2 on page 171). On small scales, peculiar velocities within galaxy groups cause a striking elongation in redshift space (sometimes dubbed 'fingers of god'), which can be used to measure the masses of galaxy clusters. On large scales there is a conspicuous flattening of the galaxy distribution, with considerably higher significance than in previous reports^{5,6}. The strength of this flattening measures the mass density associated with the large-scale structure: more mass generates stronger gravity and thus larger accelerations and velocities. Detailed analysis suggests that the amount of mass associated with the clustering is approximately 30% of the cosmic 'critical density', the value at which the mass of the Universe is just sufficient to eventually stop the Hubble expansion because of the backwards pull of gravity. The estimated density is consistent with a variety

of other methods, suggesting that cosmologists are finally converging on a reliable estimate of the mean mass density of the Universe. All indications point towards an infinite Universe that will expand forever.

Although cosmologists may be close to an accurate measure of the mass of the Universe, and studies currently underway^{7,8} will lead to even more precise values in the next ten years, there is still no compelling explanation for why the Universe contains such a complex mix of particles. Ordinary matter contributes about 3% to the critical mass density, whereas heavy neutrinos contribute at least 0.3% to the critical density. The mysterious 'dark matter' — presumably an as-yet undiscovered elementary particle — is thought to comprise the bulk of the mass density in the large-scale structure that Peacock and collaborators are measuring.

Last year, two CMB experiments^{9,10} provided convincing evidence that the Universe is 'flat', meaning that the density of the Universe exactly equals the critical density. The year before that, astronomers found evidence that the expansion of the Universe is actually accelerating, as suggested by teams studying distant supernovae^{11–13}. Both of these observations can be explained if the Universe today is dominated by smoothly distributed 'dark

energy', which constitutes the other 70% of the critical density and is thought to cause 'cosmic repulsion' on large scales. This dark energy might be Einstein's cosmological constant, or it might be the manifestation of an active quantum field of mysterious origin¹⁴.

As additional pieces of the puzzle fall into place, our picture of Big Bang cosmology has become ever more bizarre. A unifying principle is clearly needed to explain the many disparate components of the Universe we have so far discovered. The quest for such unification is likely to keep us busy for decades to come.

Marc Davis is in the Department of Astronomy and Physics, University of California, Berkeley, California 94720, USA.

e-mail: marc@deep.berkeley.edu

1. Peacock, J. A. *et al. Nature* **410**, 169–173 (2001).
2. da Costa, L. *et al. Astrophys. J.* **424**, L1–L4 (1994).
3. Penzias, A. A. & Wilson, R. *Astrophys. J.* **142**, 419–421 (1965).
4. Bennett, C. L. *et al. Astrophys. J.* **464**, L1–L4 (1996).
5. Cole, S., Fisher, K. & Weinberg, D. *Mon. Not. R. Astron. Soc.* **275**, 515–526 (1995).
6. Taylor, A. N. *et al. Mon. Not. R. Astron. Soc.* (in the press).
7. Sloan Digital Sky Survey, <http://www.sdss.org/sdss.html>
8. Microwave Anisotropy Probe, <http://map.gsfc.nasa.gov/>
9. de Bernardis, P. *et al. Nature* **404**, 955–959 (2000).
10. Hanany, S. *et al. Astrophys. J.* **545**, L5–L9 (2000).
11. Schmidt, B. *et al. Astrophys. J.* **507**, 46–63 (1998).
12. Goobar, A. *et al. Physica Scripta* **85**, 47–58 (2000).
13. Perlmutter, S. *et al. Nature* **391**, 51–54 (1998).
14. Wang, L. *et al. Astrophys. J.* **530**, 17–35 (2000).

Ageing

Yeast longevity gene goes public

David Gems

Changes in gene silencing throughout life might be a general phenomenon underlying ageing and longevity: this mechanism is at work in yeast and, as new work suggests, nematode worms.

Would it make sense to study the weed *Arabidopsis thaliana* to understand human consciousness? How about trying to comprehend human ageing by looking at the budding capacity of a fungus? Research using 'model' experimental organisms often involves gambling on the universality of biological characteristics. Arguably, using budding yeast (*Saccharomyces cerevisiae*) to investigate the genetic determinants of ageing, and consequently longevity, seems almost absurdly optimistic. Ageing yeast do not develop grey hair or poor eyesight, or start complaining about young people today, or have strokes. In fact, it is not even clear that they age at all, and when researchers talk of yeast 'lifespan', what they really mean is the number of times a yeast mother cell can reproduce by producing a bud.

Mutations in several genes can increase the lifespan of yeast. That may be good news for this unicellular fungus, but does it mean anything for multicellular creatures such as ourselves? The answer may be 'yes'. On page 227 of this issue¹, Tissenbaum and Guarente

reveal that *sir-2.1*, a relative of a yeast gene that controls lifespan, also controls longevity in an animal — the tiny nematode worm *Caenorhabditis elegans*. After more than a decade of guessing that studies of *S. cerevisiae* might teach us about general mechanisms of ageing, the gamble has paid off.

In *C. elegans*, *S. cerevisiae*², the fruitfly *Drosophila melanogaster*³, and even mice^{4,5}, there are many genes that, when mutated, increase longevity. In many cases, the proteins encoded by such genes have equivalents in higher animals. For example, the adult lifespan of *C. elegans* can be tripled as a result of reduced activity of a signalling pathway resembling that which responds to insulin or insulin-like growth factor-I (IGF-I) in mammals².

But of course what one really wants to know is whether such genes control ageing (and so longevity) in all animals. To put it more broadly, has the biology of ageing been conserved throughout evolution? Clearly, most types of animal grow old and die (although there may be exceptions, such as

Hydra)⁶. But it does not necessarily follow that ageing involves the same processes in all species. The widespread occurrence of ageing could be attributed to the common evolutionary conditions that give rise to it, in particular the fact that there is little selective pressure in old age on the effects of a gene that have little impact on reproductive success⁷. It is possible that, in different species, the underlying mechanisms of ageing are different.

The gerontologist George Martin has drawn a distinction between 'private' ageing mechanisms, which are unique to one species or animal group, and 'public' ones, such as oxidative damage to DNA, which are likely to be common to all species⁸. To find out whether a given gene, identified as a longevity determinant in one species, affects public or private mechanisms of ageing, one can test whether its equivalent in a very different species has a similar function. This is what Tissenbaum and Guarente¹ set out to do with the *sir-2.1* gene.

The yeast Sir2 protein controls gene silencing, by altering the structure of chromatin — the complex of DNA, RNA and protein that makes up chromosomes. Sir2 silences genes at various DNA sites, such as ribosomal DNA, repeated stretches of sequence that encode ribosomal RNA. Ribosomal DNA is important to yeast ageing: during cell division, circles of ribosomal DNA that are separate from the main chromosomes are generated; these circles reduce yeast lifespan⁹. Experimental overexpression of the *SIR2* gene results in reduced formation of these extrachromosomal DNA circles and extended lifespan¹⁰. But the accumulation of these circles has not yet been detected in higher organisms. So, like ageing-related accumulation of circles derived from mitochondrial DNA in the fungus *Podospora anserina*¹¹, the formation of the extrachromosomal ribosomal DNA circles seems likely to represent a 'private' mechanism of fungal ageing.

But *SIR2* is also involved in another ageing mechanism in yeast — a mechanism that is likely to be 'public'. In many animal species (possibly including primates), reduction of caloric intake extends lifespan. In yeast, this requires functional *SIR2* (ref. 12), and is independent of the formation of extrachromosomal ribosomal DNA circles. So it seemed worth testing whether *SIR2* might affect lifespan in other organisms, too. And indeed, Tissenbaum and Guarente¹ show that overexpression of *sir-2.1*, the *C. elegans* counterpart, increases mean lifespan in worms by up to 50%.

The authors then ask whether *sir-2.1* acts through one of the known pathways involved in longevity in *C. elegans* — the insulin/IGF pathway². This pathway includes the *daf-2*, *age-1* and *pdk-1* genes. These genes, perhaps in response to the presence of food¹³, inhibit *daf-16*, which encodes a transcription factor.

Mutation of *daf-2*, *age-1* or *pdh-1* results in the activation of *daf-16* and increased lifespan. Mutation of *daf-16* blocks this extension in lifespan.

Tissenbaum and Guarente find that, first, the life extension resulting from overexpression of *sir-2.1* is prevented by mutation of *daf-16*. This places *daf-16* downstream of *sir-2.1* in the insulin/IGF pathway. Second, the life-extending effects of overexpressing *sir-2.1* and mutating *daf-2* are not additive, again placing these genes in the same pathway. Further results confirm this. It seems that *sir-2.1* antagonizes the pathway: whereas *daf-2* acts to inhibit the downstream *daf-16* gene, and can no longer do so when mutated, *sir-2.1* works to activate *daf-16*, and can do so more strongly when overexpressed. The authors suggest that the silencing of genes upstream of *daf-16* by *sir-2.1* is involved in the response of the insulin/IGF pathway to caloric restriction.

Of course it remains to be shown that *sir-2.1* is involved in modifying chromatin in worms. Furthermore, a previous study involving *C. elegans* suggested that caloric restriction and insulin/IGF signalling are independent of each other¹⁴. Nonetheless, the idea of the regulation of animal longevity through changes in chromatin structure has appeal. In development, for example in pattern formation in *Drosophila*, changes in

chromatin structure have a major role. Perhaps what these latest findings¹ are telling us is that the passage through adulthood involves changes in gene regulation that are analogous to those occurring during development¹⁵, and are regulated through changes in chromatin structure. At the very least, it seems that some genetic determinants of longevity and ageing are conserved across animal groups — a fact that will encourage those studying ageing in model organisms. ■

David Gems is in the Department of Biology, University College London, 4 Stephenson Way, London NW1 2HE, UK.

e-mail: david.gems@ucl.ac.uk

1. Tissenbaum, H. A. & Guarente, L. *Nature* **410**, 227–230 (2001).
2. Guarente, L. & Kenyon, C. *Nature* **408**, 255–262 (2000).
3. Lin, Y.-J., Seroude, L. & Benzer, S. *Science* **282**, 943–946 (1998).
4. Brown-Borg, H. M., Borg, K. E., Meliska, C. J. & Bartke, A. *Nature* **384**, 33 (1996).
5. Coschigano, K., Clemmons, D., Bellush, L. & Kopchick, J. *Endocrinology* **141**, 2608–2613 (2000).
6. Martinez, D. E. *Exp. Gerontol.* **33**, 217–225 (1998).
7. Williams, G. C. *Evolution* **11**, 398–411 (1957).
8. Martin, G. M., Austad, S. N. & Johnson, T. E. *Nature Genet.* **13**, 25–34 (1996).
9. Sinclair, D. & Guarente, L. *Cell* **91**, 1033–1042 (1997).
10. Kaerberlein, M., McVey, M. & Guarente, L. *Genes Dev.* **13**, 2570–2580 (1999).
11. Jamet-Viennery, C., Begel, O. & Belcour, L. *Cell* **21**, 189–194 (1980).
12. Lin, S., Defossez, P. & Guarente, L. *Science* **289**, 2126–2128 (2000).
13. Apfeld, J. & Kenyon, C. *Nature* **402**, 804–809 (1999).
14. Lakowski, B. & Hekimi, S. *Proc. Natl Acad. Sci. USA* **95**, 13091–13096 (1998).
15. Gems, D. *J. Anat.* **197**, 521–528 (2000).

Condensed-matter physics

First glimpse of the orbiton

Philip B. Allen and Vasili Perebeinos

Crystals are host to a variety of disturbances, which show up as waves or particles. Physicists think they may have seen a new type of disturbance called the orbiton.

On page 180 of this issue, Saitoh *et al.*¹ present evidence for the observation of a new particle in a solid. Although predicted to exist, this first sighting of the orbiton should allow it to join the list of 'particle excitations' found in solids, such as electrons, holes, phonons, magnons, excitons and plasmons. Working at low temperatures, the authors measured the frequency shift of the laser light scattered from a particularly perfect crystal of LaMnO₃. This ceramic is the parent compound of a group of materials whose electrical properties switch under a magnetic field.

Modern condensed-matter theory asserts that a solid crystal is actually a gas of weakly interacting 'quasi-particles'. The prefix quasi is unnecessary. These particles are as real as the more familiar photons, electrons and protons. Particles in physics are generally connected with a property of the ground state known as 'broken symmetry'. In solids, the crystalline ground state is the lowest energy state.

Symmetry breaking is a common occurrence. For example, a pencil balanced on its tip has rotational symmetry — it looks the same from all sides — but when it falls it lands in one direction, breaking the symmetry. The fallen pencil also creates 'order' because it picks one direction in a previously unordered world. In physics, breaking symmetry and creating order go hand-in-hand. Perfect symmetry is an ideal vacuum with no order at all, but particles in such a vacuum are very dull. Fortunately, the real world is less symmetric, allowing more interesting particles to emerge.

In condensed-matter physics, symmetry breaking also leads to new particles. Consider a crystal of metallic iron. Each atom is surrounded by a lattice of neighbouring atoms with the symmetry of a cube. Left, right, up, down, front and back all look the same to the atom, until a compass needle reveals a magnetic field, pointing along one of the crystal axes. In simple terms, each iron atom has its



100 YEARS AGO

Commander R. F. Scott, R. N., in naval charge of the British Antarctic Expedition, has stated to a representative of Reuter's Agency that the preparations for the British Antarctic Expedition are now practically complete. The *Discovery*, the expedition's ship, will be launched on March 23, and, after she has been handed over by the contractors, will come round to London, where her equipment and provisions will be put aboard. The *Discovery* has been built on whaler lines, only with greatly increased strength to withstand ice pressure. She is 171 feet long, and 34½ feet beam, and has 1500 tons displacement. She will have auxiliary steam, and is fitted with engines of the latest type. In her construction the lines of the *Fram*, though carefully studied, have not been adopted, as Nansen's ship would have been ill-adapted for the heavy seas the *Discovery* will have to encounter. The expedition will leave London in July or August, and will proceed to Melbourne, reaching there in November. From *Nature* 7 March 1901.

50 YEARS AGO

The annual report of the Chicago Natural History Museum for 1949 (pp. 140; Chicago: Chicago Natural History Museum, 1950; 1 dollar) gives some interesting facts concerning the activities of its Department of Public Relations. For an entire year, the Museum was brought to the notice of every person who looked up a telephone number in the "Red Book, Chicago Classified Telephone Directory"; for, through the courtesy of the publishers, the name of the Museum and a picture in colours of the exterior appeared on the front cover, while inside in a prominent page-one position was the story of the Museum. Considering the large number who consult a telephone directory in the course of a year, the cumulative effect of such publicity must be of a high order ... The museum continued to be represented each Saturday throughout the year by a series of stories on the "Children's Corner" radio programme. Special features, including television, were given on several of the broadcasting systems. Other publicity included the distribution of thousands of folders describing exhibits. Lecture courses for adults and children were advertised by posters displayed in railway stations and on suburban trains. Some of the methods adopted by the American museums may seem unusual; but all are worthy of consideration.

From *Nature* 10 March 1951.

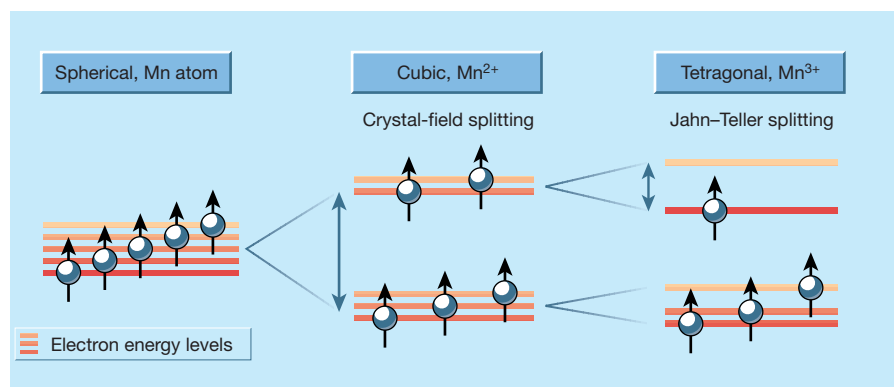


Figure 1 How electron energy levels are affected by their environment. Energy level diagrams for d orbital electrons (upward arrows indicate direction of spin) of a manganese (Mn) atom in free space (left, spherical environment), in a cubic environment (middle, such as divalent MnO), and in a tetragonal environment (right, such as trivalent LaMnO_3). Saitoh *et al.*¹ now claim to have seen orbitons — a collective oscillation of partially occupied electron subshells — in a pure crystal of LaMnO_3 .

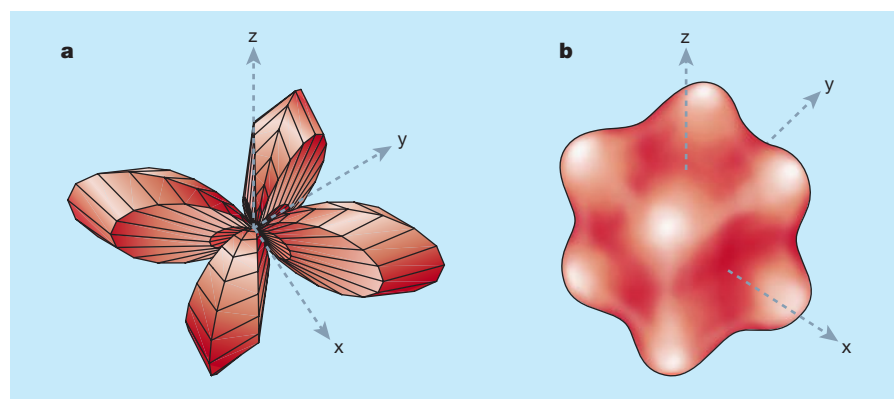


Figure 2 The shape of electron orbitals. a, The charge density of a single d orbital electron and b, the total charge density of a filled subshell with three d orbital electrons. The filled subshell has cubic symmetry. If both subshells were filled, the total charge density would look like a sphere, indistinguishable from the spherical symmetry of free space.

own 'magnetic spin'. At temperatures above 770°C , the spins turn rapidly and randomly; magnetism is not detected. But at lower temperatures, the spins spontaneously align, randomly choosing one of the six equivalent directions. As with the pencil, the rotational symmetry has been broken.

This symmetry breaking allows new disturbances in the solid, which physicists recognize as waves, or (with the wave-particle duality of quantum theory) as particles. For example, if the spin of one iron atom is turned away slightly from the overall direction of the magnetization, the neighbouring spins exert a restoring force. Rotational inertia prevents these spins from responding instantly, and wave propagation occurs when restoring forces encounter inertia. Therefore, spin disturbances propagate as spin waves, or particles called magnons.

But not all particles are straightforward to detect. Photons in light waves can be measured directly when they enter a photomultiplier tube. But magnons in iron can only be detected indirectly, by scattering a beam of photons or neutrons from the iron crystal. Inside the crystal a photon or neutron may interact with a magnon, causing

detectable changes in the properties of the scattered particles. This is how Saitoh *et al.*¹ set about looking for orbitons in LaMnO_3 .

The tetragonal crystal structure of the ceramic contains magnetic manganese ions (Mn^{3+}). In addition to magnetic spins, which become ordered at -123°C (ref. 2), the Mn^{3+} ion has a partly filled electron shell, which provides a new way to break symmetry and create 'orbital order'. Orbitals are the allowed energy states for electrons. When occupied by an electron, they are visualized as clouds of electron charge. Symmetry breaking in LaMnO_3 orbitals happens at about 475°C . Below this temperature, LaMnO_3 has orbital order³, and orbital waves, or their corresponding orbitons, may exist.

To understand orbital order we need to review a little chemistry. Stable chemical species have filled electron shells. Only two electrons, one of each 'spin orientation', can be associated with each orbital, but each shell contains a set of orbitals. In atoms the shells are labelled $1s$, $2s$, $2p$, $3s$, $3p$, $3d$, $4s$ and so on. Argon, for example, is inert. It has 18 electrons filling the shells $1s$ through to $3p$. The s shell has one orbital, which is spherically symmetric. The p shell contains three

orbitals. The filled p shell is spherically symmetric, but the individual p orbitals are elongated in three perpendicular directions (a bit like dumbbells).

The manganese atom, on the other hand, has seven electrons more than argon, and so has two electrons (of opposite spin) in a filled $4s$ shell and five electrons with the same spins in a half-filled $3d$ shell. Half filling is the next best thing to complete filling. Each individual d orbital is asymmetric, but when all the orbitals are occupied by one electron, the overall charge density is spherically symmetric. Figure 1 shows energy level diagrams for the five d orbitals of atomic manganese in different environments. In a cubic solid, the Mn^{2+} ion (as in MnO) loses its two $4s$ electrons but keeps the half-filled $3d$ shell. Because spherical symmetry is broken by the cubic crystal lattice, the $3d$ shell is split into two subshells. Neither subshell alone has full spherical symmetry, but they still have cubic symmetry (Fig. 2).

Now consider the Mn^{3+} ion in the environment of LaMnO_3 . The $3d$ shell now has only four electrons (Fig. 1). At high temperatures the Mn^{3+} ion has cubic symmetry: the lower subshell is filled and the upper subshell is half-filled. To reduce its energy, this ion would prefer a lower symmetry in which the upper subshell splits in two. So below 475°C , LaMnO_3 stretches along one axis of the crystal to become tetragonal. The occupied d orbital electrons now have a tetragonal shape, and the crystal has orbital order (also called cooperative Jahn-Teller order).

In LaMnO_3 , the tetragonal shape of the occupied $3d$ orbital in the upper subshell can be deformed. There is a restoring force discouraging such deformation, and inertia to slow the process. Therefore, orbital wave propagation is expected — a collective oscillation of partially occupied electron subshells. This is the orbiton particle that Saitoh *et al.*¹ claim to have seen.

In the experiment, Saitoh and colleagues picked out a component of the scattered laser light whose frequency shift corresponds, they believe, to excitation of an orbiton. With increasing temperature, this component weakens and completely disappears above 475°C , the temperature at which cubic symmetry is restored and orbital order is lost. Their identification of the orbiton rests on a theoretical calculation that agrees nicely with the low-temperature data. But the case is not airtight. For simplicity, the theory⁴ omits the coupling of orbitons to phonons (vibrations of the crystal lattice)^{5,6}. It needs to be shown that good agreement remains when this effect is included, both at low and at room temperature.

Orbitons should play a role in the thermal properties of orbitally ordered solids such as LaMnO_3 , so they could be looked for in measurements of heat capacity and heat transport in crystals. Their effect on magnon

and phonon spectra should also be calculated and measured. One might expect orbitons to occur in other orbitally ordered materials, such as high-temperature superconductors. Indeed, the partially empty d orbital shell of Cu^{2+} in copper oxide superconductors can have orbital excitations. Over a decade ago, these excitations were used by Weber⁷ in a theory of the superconducting coupling between electrons that allows them to flow without resistance. Perhaps he should revive his theory now that the orbiton's existence is more established. ■

Philip B. Allen and Vasili Perebeinos are in the Department of Physics and Astronomy, State University of New York, Stony Brook, New York 11794-3800, USA.

e-mail: allen@felix.physics.sunysb.edu

1. Saitoh, E. *et al.* *Nature* **410**, 180–183 (2001).
2. Moussa, F. *et al.* *Phys. Rev. B* **54**, 15149–15155 (1996).
3. Murakami, Y. *et al.* *Phys. Rev. Lett.* **81**, 582–585 (1998).
4. Okamoto, S., Ishihara, S. & Maekawa, S. *Phys. Rev. B* **61**, 14647–14655 (2000).
5. Hotta T., Malvezzi, A. L. & Dagotto, E. *Phys. Rev. B* **62**, 9432–9452 (2000).
6. Allen, P. B. & Perebeinos, V. V. *Phys. Rev. Lett.* **83**, 4828–4831 (1999).
7. Weber, W. Z. *Phys. B* **70**, 323–329 (1988).

Evolutionary biology

What's in a baboon's behind?

R. I. M. Dunbar

It is often thought that sexual 'ornaments', such as the swellings that adorn ovulating female baboons, are signalling something about fertility — but what? Long-term studies of wild baboons provide an answer.

The Victorians were so embarrassed by the large swellings that occasionally adorn the bottoms of some primate species — notably baboons and chimpanzees — that many zoos insisted that females in this condition be kept out of public view. The swellings, which occur in about 10% of all primate species, can be so large and grotesque that females cannot sit comfortably. Less squeamish primatologists of subsequent eras remained puzzled by this all but unique phenomenon. The fact that swellings coincide with ovulation made it obvious that females must be signalling something about their fertility, but exactly what remained elusive. On page 204 of this issue¹, Domb and Pagel show that females are in fact signalling their own genetic quality.

Sexual swellings are found only among Old World monkeys and apes. Detailed analyses of the distribution of swellings indicate that they have evolved independently in three quite separate lineages. Typically they are associated with the evolution of social groups containing several males and several females, where mating is promiscuous and females have greater choice in mating partners^{2–5}. Swellings are seen, for example, in chimpanzees, whose mating system is close to a sexual free-for-all. But — as readers have no doubt noticed — they do not occur in the chimpanzees' closest relatives (our own species), whose mating system is typically, though by no means exclusively, focused on monogamous couples.

Five explanations for the evolution of sexual swellings have been offered^{4,6}. First, females want to attract males to join their groups, so as to provide the females and their young with an effective defence against predators. Second, females are in effect try-

ing to provoke males into competing among themselves for opportunities to mate, thereby enabling the females to mate with the 'best' males in the group (a form of indirect female choice). Third, females are encouraging promiscuous mating so as to confuse paternity. This would mean that the risk of killing their own offspring deters males from infanticide — a problem that is particularly serious for primates because they have long intervals between births⁷. Fourth, given that the timing of ovulation is itself unpredictable, females want to persuade individual males to stay with them. This would assure the males of paternity, and provide the females with a bodyguard so as to reduce the levels of harassment that they would otherwise suffer in multimale groups (this is the 'hired-gun' hypothesis). And fifth, swelling size is an indicator of a female's genetic quality, enabling her to attract the best males to mate with her and so maximize the genetic quality of her offspring⁸.

One of the problems that evolutionary biologists have encountered in trying to test these hypotheses is that they often make similar predictions. As a result, comparative data drawn from across the primate species are often compatible with several hypotheses; indeed, there is evidence to support all five hypotheses⁴. In such cases, to choose between competing theories we need to explore the mechanisms involved in more detail, rather than just look at the superficial phenomenon.

Domb and Pagel¹ provide the first uncontroversial mechanistic evidence that the size of sexual swellings is correlated with a male's willingness to compete for individual ovulating females, and with indices of a female's own fitness. The authors were able to reach this conclusion only because they used long-

term demographic data from the baboon population that (together with Jane Goodall's more famous chimpanzees) inhabits the Gombe National Park, Tanzania. Both the chimp and the baboon populations have been under continuous study for more than 30 years.

Domb and Pagel show that the females with the largest swellings had attained sexual maturity earlier (and so had longer reproductive lifespans) than those with smaller swellings. The females with larger swellings also produced more offspring, as well as more surviving offspring, over their lifetime. Males competed more heavily, and incurred more injuries from fighting, for females with larger swellings. As males are willing to do this, it follows that they must be using swelling size as an index of a female's heritable reproductive quality. These findings thus explain why there are such consistent differences between females in the size and quality of their swellings. But it may not explain why sexual swellings evolved in the first place — it is possible that females have exploited for this new purpose a signal that first evolved for one of the other reasons.

Evolutionary biologists have been much exercised of late as to whether such signals of quality have to be honest, by which they mean costly to produce. Although there is some evidence to suggest that this need not always be so, most theoretical models of signal evolution agree that when cues are cheap to produce they are too susceptible to cheating, and so are evolutionarily unstable. Herein, perhaps, lies the weak link in Domb and Pagel's argument — they are able to provide only circumstantial evidence that swellings are costly for females to produce. Their evidence is that female body weight increases by 14 to 25% when swellings are at their maximum size. It is unlikely that the swellings are in themselves energetically expensive, because they are produced simply by the selective retention of extracellular water. But it is possible that they make females less fleet — and thus at greater risk of being caught by a predator — and that carrying the added weight increases the energetic costs of travel. If either of these could be demonstrated, the case would be satisfyingly closed. ■

R. I. M. Dunbar is in the Population and Evolutionary Biology Division, School of Biological Sciences, Nicholson Building, University of Liverpool, Liverpool L69 3GS, UK.
e-mail: rimd@liv.ac.uk

1. Domb, L. G. & Pagel, M. *Nature* **410**, 204–206 (2001).
2. Hrdy, S. B. & Nicholson, N. in *Primate Societies* (eds Smuts, B. B., Cheney, D. L., Seyfarth, R. M., Wrangham, R. W. & Struhsaker, T. T.) 370–384 (Chicago Univ. Press, 1987).
3. Sillén-Tullberg, B. & Møller, A. P. *Am. Nat.* **141**, 1–25 (1993).
4. Nunn, C. L. *Anim. Behav.* **58**, 229–246 (1999).
5. Dixon, A. F. *Primate Sexuality* (Oxford Univ. Press, 1998).
6. Manson, J. H. *Curr. Anthropol.* **38**, 353–374 (1997).
7. van Schaik, C. P. & Kappeler, P. M. *Proc. R. Soc. Lond. B* **264**, 1687–1694 (1997).
8. Pagel, M. *Anim. Behav.* **47**, 1333–1341 (1994).

Materials science

Salting the surface

Marshall Stoneham

When one type of material is grown on a very different one, the electrostatic interactions between them become important. These interactions offer new ways to control structure on the nanoscale.

In nanotechnology, control is all-important. It is easy to produce interesting structures by chance, but much harder to build or replicate a chosen structure by design. So it makes sense to look for general principles that can guide construction even when the requirements become complex. Many of these guiding rules are extremely specialized, relying on very particular bonding behaviours or atomic geometries. So it is a step forward when a new general principle is suggested. An example appears in *Physical Review Letters*, in a recent analysis by Repp *et al.*¹ of the interface between a salt and a metal. Their work casts a new slant on electrostatic interactions between surfaces, which could provide a way of controlling structure on the nanoscale.

Most useful designs involve contact between materials that are very different in character. For example, interfaces between conductive metals and insulating ionic crystals, such as oxides and halides, are common. The oxide on metals can be helpful (as a protective oxide on aluminium beverage cans) or harmful (such as rust or corrosion products). At a smaller scale, one can use an ultra-thin gate oxide on silicon in microelectronics, or metal particles on oxide substrates for catalysis. Understanding the nature of these interfaces at the nanoscale may allow greater control over the final product. With interfacial chemistry, a few general rules stem from practical experience. The most useful of these can be applied to real systems, even when such systems are too complex to be described by state-of-the-art basic science.

The most familiar of these key ideas come from crystallography, the study of crystal structures. Here, epitaxy — the geometry of fitting two crystal lattices together — is important for controlling the interface between the lattices. Then there are ideas that use macroscopic analogies with surface tension, produced when liquids wet surfaces. Another group of ideas concerns the elasticity of interfaces — for example, how a material can be deformed by structures on its surface, leading to strains that drive self-organization and restructuring.

The work of Repp *et al.*¹ emphasizes another set of ideas relating to electrostatics, in particular, the way one medium (a relatively unreactive metal) is affected by electric charges in the other (an ionic salt). We know that if a conductive material, such as copper, is in contact with an ionic material, for exam-

ple sodium chloride, the positions and charges of the ions redistribute the charge in the metal. Of course, the various interactions often influence each other. Elastic strain might interfere with epitaxy, leading to misfit dislocations, or to intermediate layers of different structure. The relative importance of the different interactions largely depends on how thick the two materials are. A very thin layer can adapt to a substrate in ways that do not work for thicker materials.

Repp and colleagues¹ have investigated the growth of a thin NaCl film on a Cu surface. Crucially, the Cu surface they have studied is not flat but has regular steps like corrugated iron. The authors used scanning tunnelling microscopy (STM) to determine the location of the sodium and chloride ions on the metal surface. The STM data show an unexpected result: the chlorine ions sit on top of the step edges, whereas the sodium ions appear to lie between the steps (Fig. 1). Such a pattern indicates a localized binding between NaCl and the underlying Cu surface. This is surprising given that Cu and NaCl are largely unreactive together.

To explain their results, Repp *et al.* revisit some old ideas involving electrostatics². The role of electrostatic interactions in metal–ionic interfaces has long been recognized³. At their simplest, ions (Na^+ , Cl^-) polarize the metal to create an attractive interaction: the ions interact with their ‘image’, or opposite, charges in the metal (so approaches like this are termed ‘image-charge models’). This model suggests that, if there is no chemical interaction, the attractive interaction might depend weakly on the

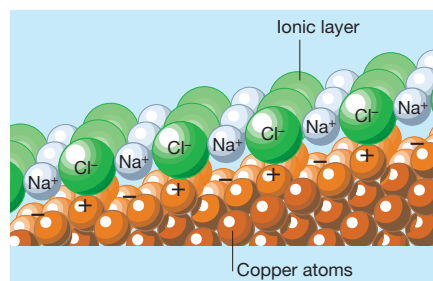


Figure 1 Charge modulation at interfaces. Repp *et al.*¹ observe an unusual pattern when they grow a salt (NaCl) on a stepped metal surface (Cu): the Cl ions are always located on top of the steps. This binding pattern can be explained by the Smolouchowski effect², which leads to a charge-corrugated Cu surface that forms a strong electrostatic bond especially with Cl ions.

metal, yet would depend significantly on the charges in the ionic compound. Changing the state of the ionic charges, perhaps by shining a light on the interface, could alter the interaction. The image-charge picture has become influential in studies of wetting by liquid metals, metal oxidation, catalysts supported by metals, and even, it has been suggested, in explaining how one can improve the grip of car tyres on ice with electricity.

Repp and colleagues argue that the original image-charge picture works best for very flat surfaces. But the stepped Cu surface is more complex. They note that the surface charge has an inbuilt structure in such situations (the Smolouchowski effect², first recognized in 1941). The gas of electrons inside the metal does its best to neutralize the charge of the positive ion cores. But for more complicated surfaces, with unevenly distributed surface atoms, certain ion cores are neutralized more effectively than others. On stepped and kinked surfaces, one would expect to see specific regions of positive or negative charge. When an ionic material grows on that surface, the favoured structures will have positive ions near negatively charged regions. And where the metal surface is positive, like the step edges on the Cu surface studied by Repp *et al.*, negative ions should be found, as was seen in the STM experiments.

Does this interpretation agree with earlier work? Most of the detailed theory exists for the reverse situation, where, for example, layers of metal (Ag) are grown on a metal oxide (MgO). The precise charge redistributions vary between relatively flat and more complex MgO surfaces⁴. There is no sign of the features discovered by Repp and colleagues. This may be because the ion charges are larger, swamping the Smolouchowski effect, or because the Ag layers were too thin. Certainly, other electrostatic features should be considered. Interactions between the Na and Cl ions could dominate the resulting surface structure. And more subtle effects are possible if there are additional kinks on the surface steps of NaCl⁵.

Nonetheless, the ideas of Repp *et al.*¹ suggest new ways of looking at electrostatic interactions and of explaining systematically what happens when one type of material is grown on a very different one. The idea that charge modulation can complement the more familiar ideas of crystallography, and suggest new ways to control the structures of thin films.

Marshall Stoneham is at the Centre for Materials Research, Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK.

e-mail: ucapams@ucl.ac.uk

1. Repp, J., Fölsch, S., Meyer, G. & Rieder, K.-H. *Phys. Rev. Lett.* **86**, 252–255 (2000).
2. Smolouchowski, R. *Phys. Rev.* **60**, 661–663 (1941).
3. Stoneham, A. M. & Tasker, P. W. *J. Phys. C* **18**, L543–L548 (1985).
4. Zhukovskii, Y. F. *et al. Surf. Sci.* **441**, 373–383 (1999).
5. Seitz, F. *Rev. Mod. Phys.* **26**, 7–94 (1954).

Prions

The shape of a species barrier

Susan W. Liebman

Intriguing *in vitro* results highlight the possibility that conformational differences between prion proteins with the same sequence contribute to the barrier that limits prion propagation between species.

The recent epidemics of 'mad cow' disease have put the fatal, transmissible spongiform encephalopathies — the so-called prion diseases — firmly in the public eye. It seems likely that the infectious agent in these mammalian diseases is the PrP protein, in the absence of traditional genetic material (nucleic acid). It is also likely that the infectious capability of the PrP protein depends upon the shape into which it is folded: when some of this protein is in its disease-causing ('prion') conformation, it converts normal PrP into that form, too¹.

Prion diseases have two particularly intriguing features. First, clinically different prion diseases, called strains, can occur within one animal species. It appears that, although the *PrP* gene does not differ between individuals of any one species, the shape of the protein it encodes can differ, and different conformations are responsible for different disease strains². The second feature is the 'species barrier' that limits the spread of infection between species. Factors influencing this barrier are of particular interest

because transmissible spongiform encephalopathies are occasionally transmitted from cattle to humans. On page 223 of this issue³, Chien and Weissman investigate and relate these two features by studying a prion protein from the budding yeasts *Candida albicans* and *Saccharomyces cerevisiae*.

Recent experiments have established the existence of other proteins, distinct from PrP, that control the inheritance of certain traits, and the term 'prion' has now been expanded to include them⁴. A yeast protein involved in propagating the so-called $[PSI^+]$ (ref. 5) trait is one such addition. $[PSI^+]$ yeast cells carry a prion form of a protein called Sup35, which is involved in terminating protein translation — the production of proteins from messenger RNA. In $[PSI^+]$ cells, much of the Sup35 is in the prion shape and forms aggregates, so is not available to terminate translation. The result can be translational read-through — that is, translation does not always stop even when the cellular translation machinery reaches a 'stop' codon in mRNA.

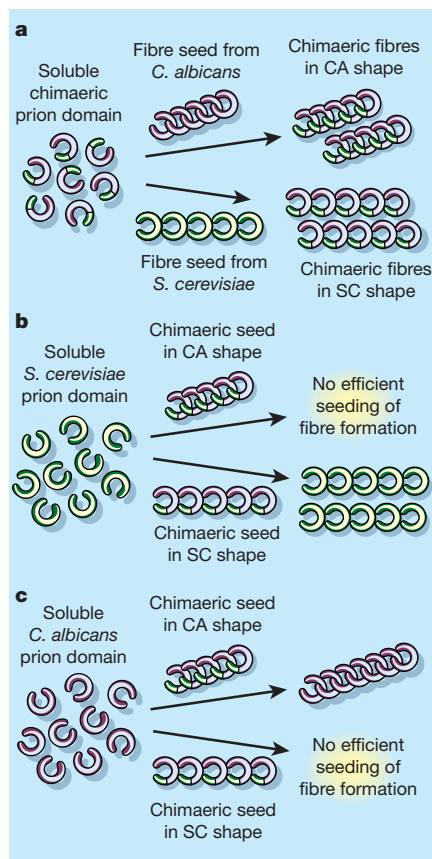


Figure 1 Model showing the involvement of fibre conformations in the species barrier between the yeasts *Candida albicans* and *Saccharomyces cerevisiae*. Chien and Weissman³ constructed a chimaeric protein containing parts of the so-called prion domains from the *C. albicans* and *S. cerevisiae* Sup35 proteins. **a**, The soluble (monomeric) chimaeric prion domain is promiscuous and able to aggregate into one of two fibre-like conformations, SC or CA, depending upon whether fibres of the wild-type *S. cerevisiae* prion domain or *C. albicans* prion domain, respectively, are used to 'seed' the reaction. Although it is not known, for simplicity the *C. albicans* and *S. cerevisiae* fibres are shown with different conformations and the chimaeric fibres are shown as having the same conformation as the inducing *C. albicans* or *S. cerevisiae* fibres. **b**, Chimaeric fibres in the SC shape, but not in the CA shape, efficiently seed the formation of fibres from the *S. cerevisiae* soluble prion domain. **c**, Chimaeric fibres in the CA but not SC shape efficiently seed the formation of fibres from the *C. albicans* soluble prion domain.

The amino-terminal region of Sup35 is called the prion domain, because it is required for the propagation of the $[PSI^+]$ trait in yeast cells. Also, when experimentally overproduced in yeast cells for a short time, that region is sufficient to induce the cells to become $[PSI^+]$. Overproduction of the prion domain induces the appearance of distinct, heritable 'strong' and 'weak' $[PSI^+]$ strains, with strong strains showing more read-through⁶. These differences might be caused by Sup35 assuming different conformations, as suggested by analogy with the mammalian PrP strains², and by the finding that yeast cells that carry the strong $[PSI^+]$ trait have less Sup35 in its unaggregated form than do weak $[PSI^+]$ cells⁷.

The purified, soluble prion domain of Sup35 forms fibres *in vitro*, but with a lag; the addition of preformed fibres ('seeds') eliminates that lag time^{8,9}. The connection between the fibres formed *in vitro* and the aggregated Sup35 found in $[PSI^+]$ yeast cells is unknown. But alterations in the prion domain have similar effects on $[PSI^+]$ phenomena *in vivo* and on the kinetics of fibre formation *in vitro*^{10–13}. Notably, two morphologically distinct shapes of fibre have been detected using electron microscopy, and these forms may correspond to different $[PSI^+]$ strains⁸. Once a fibre begins to polymerize in one conformation, that shape is maintained as the fibre grows⁸.

Can a prion form of Sup35 from one species of budding yeast convert the Sup35 from another species to the prion form? It seems not. When the prion domains of Sup35 from other yeast species are artificially expressed in *S. cerevisiae*, they can take on the prion conformation. But, as with mammalian PrP, the prion form of Sup35 from one yeast species is not efficiently transmitted *in vivo* to the Sup35 prion domain of another species^{11,14,15}. Likewise, the speed of formation of fibres from the Sup35 prion domain is enhanced only if the reaction is seeded with fibres from the same species¹¹. There is clearly a barrier between yeast species, and Chien and Weissman³ aimed to find out whether prion conformation might be involved.

The authors constructed a chimaeric prion domain, composed of portions of the prion domains from the *C. albicans* and *S. cerevisiae* Sup35 proteins. The fibres formed from the purified chimaeric molecules could take on either of two conformations, designated CA or SC, depending upon whether they were seeded with fibres produced from the 'natural' *C. albicans* or *S. cerevisiae* prion domains, respectively (Fig. 1a). Moreover, the SC chimaeric fibres could then seed the formation of fibres only from the natural *S. cerevisiae* prion domain (Fig. 1b), not from the *C. albicans* prion domain (Fig. 1c). The opposite was true for the CA fibres (Fig. 1b, c).

Chien and Weissman also engineered

S. cerevisiae cells to carry a Sup35 protein that had the chimaeric prion domain instead of the natural, *S. cerevisiae* prion domain. Cells bearing this Sup35 protein in its non-prion form could be changed to a state analogous to $[PSI^+]$, called $[X^+]$, when the natural prion domain from *C. albicans* or *S. cerevisiae* Sup35 was temporarily overproduced. Surprisingly, most of the $[X^+]$ strains induced by the *C. albicans* domain were strong, while most of those induced by the *S. cerevisiae* domain were weak or unstable. Perhaps these strong and weak *in vivo* strains correspond to the different chimaeric conformations formed *in vitro* with *C. albicans* or *S. cerevisiae* fibre seeds. *In vivo* proof of this will require the demonstration either that direct introduction of the differently shaped fibres into yeast specifically transmits the strong or weak $[X^+]$ strain, or that Sup35 aggregates isolated from yeast with the strong or weak strains have the expected differences in conformation and seeding specificity.

The authors³ have shown that, depending on the species from which the seeding fibre originates, their purified chimaeric prion domain can fold into two different conformations, themselves with distinct *in vitro* seeding specificities. The results suggest that prion conformation may hamper the spread of prion diseases between species. But, given that distinct prion strains within one species may be caused by differing prion conforma-

tions, the results also hint that those conformations may vary in their ability to cross species barriers. With their strong and weak $[X^+]$ yeast strains, Chien and Weissman have opened the door for this theory to be tested *in vivo*. It may be time to consider the disturbing possibility that certain bovine prion forms have an enhanced ability to cross the species barrier to humans.

Susan W. Liebman is in the Department of Biological Sciences, University of Illinois, Chicago, Illinois 60680, USA.

e-mail: sueL@uic.edu

1. Prusiner, S. B. *Proc. Natl Acad. Sci. USA* **95**, 13363–13368 (1998).
2. Caughey, B. & Chesebro, B. *Trends Cell Biol.* **7**, 56–62 (1997).
3. Chien, P. & Weissman, J. S. *Nature* **410**, 223–227 (2001).
4. Wickner, R. B. *Science* **264**, 566–569 (1994).
5. Cox, B. S., Tuite, M. F. & McLaughlin, C. S. *Yeast* **4**, 159–178 (1988).
6. Derkatch, I. L., Chernoff, Y. O., Kushnirov, V. V., Inge-Vechtomov, S. G. & Liebman, S. W. *Genetics* **144**, 1375–1386 (1996).
7. Zhou, P. *et al.* *EMBO J.* **18**, 1182–1191 (1999).
8. Glover, J. R. *et al.* *Cell* **89**, 811–819 (1997).
9. King, C. Y. *et al.* *Proc. Natl Acad. Sci. USA* **94**, 6618–6622 (1997).
10. DePace, A. H., Santos, A., Hillner, P. & Weissman, J. S. *Cell* **93**, 1241–1252 (1998).
11. Santos, A., Chien, P., Osherovich, L. Z. & Weissman, J. S. *Cell* **100**, 277–288 (2000).
12. Kochneva-Pervukhova, N. V. *et al.* *EMBO J.* **17**, 5805–5810 (1998).
13. Liu, J. J. & Lindquist, S. *Nature* **400**, 573–576 (1999).
14. Chernoff, Y. O. *et al.* *Mol. Microbiol.* **35**, 865–876 (2000).
15. Kushnirov, V. V., Kochneva-Pervukhova, N. V., Chechenova, M. B., Frolova, N. S. & Ter-Avanesyan, M. D. *EMBO J.* **19**, 324–331 (2000).

As a result, Little's model promised superconductivity at substantially higher temperatures (possibly above room temperature) than is allowed by the BCS phonon mechanism — but it also imposed severe constraints on the design of any such material. For example, Little's model was based on using conducting organic polymer chains modified with highly polarizable dye-like side groups. Until Schön and colleagues' experiment, the only polymeric material to show superconductivity was a crystalline inorganic polymer⁴. So although Schön *et al.* have obtained only a modest critical temperature for their polymer system, can we say that their polymer's superconductivity fits with Little's theory?

To explore this idea, we must first understand why some polymers conduct electricity. Conducting polymers, such as P3HT, have a 'conjugated' structure of alternating single and double bonds linking the carbon atoms in their backbone (Fig. 1). This bonding structure allows one electron from each carbon atom to move along the molecule and become conducting⁵. But the one-dimensional nature of the electron movement along the polymer chain makes them relatively poor conductors, especially at low temperature. This all changed 25 years ago, when it was discovered that chemically doping these materials with suitable impurities — halogens such as chlorine or alkali metals such as sodium, for example — leads to a surplus or deficit of electrons, making the polymers almost as good as metallic conductors. This finding widened the possible applications of polymer devices in electronic and optical systems, and was recognized by the Nobel Prize in Chemistry last year.

Despite their high conductivity at room-temperature, doped polymers do not really mimic metals because their conductivity is not enhanced when they are cooled. Better conductivity — and ultimately superconductivity — at lower temperatures is the trademark of a good metal. The problem for the polymers is that their electronic properties are dominated by structural disorder, which leads to charge localization at lower temperatures. This is a shame because there is evidence for strong electron–phonon interactions in polymers — favourable ingredients for BCS superconductivity⁶.

In an effort to minimize the problem of disorder, Schön and colleagues¹ used a solution-casting process to exploit the self-organizing properties of polymer films deposited on a substrate. And instead of using chemical doping, the team injected charge carriers into the system using a field-effect transistor (FET) geometry. In an FET, applying a voltage to a 'gate electrode' triggers a flow of charge along the surface of the thin film. A similar procedure had been used previously to make conducting thin films of oligothiophenes^{6–8}. A later study had shown that

Condensed-matter physics

Superconducting plastic

Denis Jérôme and Klaus Bechgaard

Polymers that can conduct electricity have been known for some time, but they have defied attempts to make them into superconductors. The answer, it turns out, is to inject them with charge — but how does it work?

Polymers make versatile plastics, but until the 1970s, they were not expected to conduct electricity. The discovery of conducting organic polymers led to fresh applications in electronics and, later, in optical devices. The main advantage of organic polymers over conventional metals and inorganic semiconductors used in electronics is their low cost. But unlike metals, they have never been made to superconduct — until now. On page 189 of this issue, Schön *et al.*¹ report superconductivity below 2.5 K in a thin film of the organic polymer poly(3-hexylthiophene) (P3HT).

Materials that can superconduct lose all their electrical resistance below a specific temperature, known as the critical temperature, T_c . For most metallic superconductors this temperature is relatively low, and the search for so-called high-temperature superconductors has been influential in the development of condensed-matter physics.

This was true, for example, in the late 1960s when the highest available T_c was raised from 18 K to 21 K by the synthesis² of a new intermetallic compound. But progress was slow, so researchers considered other systems such as polymers. This prompted W. A. Little to suggest a provocative new theory³, which offered an alternative to the popular Bardeen, Cooper and Schrieffer (BCS) picture of superconductivity in conventional metals. In the BCS theory superconductivity occurs because tiny vibrations in the underlying crystal lattice — known as phonons — help electrons to overcome their mutual repulsion and form pairs that flow without resistance.

For his model, Little replaced the phonons with an 'exciton' mechanism in which electron pairing is mediated by other electrons. These electron–electron interactions are characterized by a typical energy that is much larger than that of the phonons.

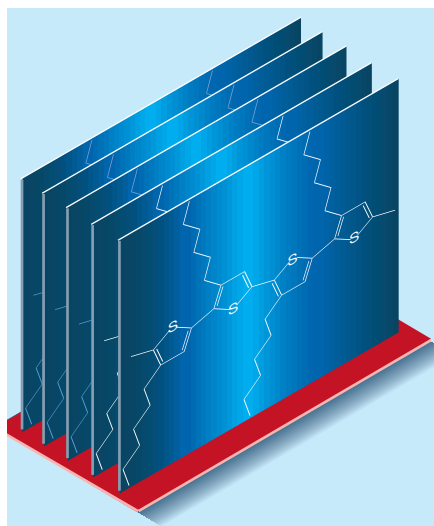


Figure 1 Polymers that can superconduct. Some organic polymers can be turned into conductors and semiconductors by chemical doping with impurities. Others can be made to superconduct by injecting them with charge carriers. In a thin film of the undoped polymer poly(3-hexylthiophene) (P3HT), conductivity occurs if charge carriers are injected. For the highest conductivity, the optimal arrangement of molecular planes of the polymer chains is perpendicular to the thin film's surface⁹. By injecting charge into a similar highly ordered film of P3HT, Schön *et al.*¹ have now observed superconductivity below 2.35 K. To inject enough charge they use a field-effect transistor geometry: source and drain electrodes are placed on top of the insulating surface (red), whereas the gate electrode is placed below. Applying a voltage to the gate electrode triggers a flow of charge along the surface of the thin film between the two other electrodes. (Redrawn from ref. 9.)

P3HT molecules can self-assemble into two-dimensional conjugated sheets, which can be oriented either parallel or perpendicular to the surface of the FET (Fig. 1)⁹. At room temperature, much higher charge mobilities are seen when these sheets are perpendicular to the FET's surface. So it appears that conduction in these polymers is not confined to the backbone of the molecule, leading to two-dimensional charge transport in the film⁹.

Studies of the P3HT film used by Schön *et al.* show that it has a nanocrystalline structure (in which ordered nanocrystals are embedded in a disordered matrix), but this does not appear to affect the two-dimensional charge transport in the film and so is not harmful for superconductivity. The density of injected charge can be monitored by the gate voltage, and the thin-film resistance exhibits a metal–insulator transition with a metallic-like temperature dependence. Increasing the charge density leads to a zero resistance state below 2.35 K, which the authors attribute to superconductivity, owing to its response to an external magnetic field. The suppression of the zero resistance

state by a high magnetic field is a defining feature of superconductivity.

Why has no one observed superconductivity in polymer films before? To achieve this experimental *tour de force*, Schön *et al.* overcame two major technical problems: the growth from solution of thin P3HT films with large nanocrystalline regions, and the tolerance of a high gate-voltage by the thin film at low temperatures. The FET gate-voltage had to be high enough to inject one charge carrier for every five polymer chains. This report will surely be the first of many investigations with the FET geometry and deposited nanocrystalline films.

Schön *et al.* also chose their polymer wisely. The arrangement of the molecular chains of P3HT is similar to the packing of donor molecules in the conducting salts of tetramethyltetraselenafulvalene (TMTSF), which were the first organic conductors to show bulk superconductivity¹⁰. But the one-dimensional character of TMTSF salts is still an important feature and possibly governs the nature of the superconducting pairing.

There is no evidence for one-dimensional transport in the polymer films studied by Schön *et al.*, and the nature of the pairing mechanism remains to be found. Strong hints in favour of a conventional explanation in terms of phonon-mediated superconductivity come from a similar study of the organic molecules pentacene, tetracene and anthracene¹¹. But it would not be altogether surprising if electron–electron interactions became important in the two-dimensional electron gas created in the FET studies.

Even if Little's model for superconductivity does not apply to these new results, we are probably witnessing a healthy revival of organic solid-state physics, which moved out of the mainstream of condensed-matter physics in the 1960s. The work of Schön *et al.* emphasizes that interdisciplinary work, involving both synthetic chemistry and condensed-matter physics, will advance the frontiers of both fields. ■

Denis Jérôme is at the Laboratoire de Physique des Solides, Université Paris-Sud, 91405 Orsay, France.
e-mail: jerome@lps.u-psud.fr

Klaus Bechgaard is at Risoe National Laboratory, 4000 Roskilde, Denmark.
e-mail: klaus.bechgaard@risoe.dk

- Schön, J. H. *et al.* *Nature* **410**, 189–192 (2001).
- Matthias, B. T. *Physics Today* August, 23–28 (1971).
- Little, W. A. *J. Polymer. Sci. Part C* **29**, 17–26 (1970).
- Greene, R. L., Street, G. B. & Suter, L. J. *Phys. Rev. Lett.* **34**, 577–579 (1975).
- McGehee, M. D., Miller, E. K., Moses, D. & Heeger, H. J. in *Advances in Synthetic Metals* (eds Bernier, P., Lefrant, S. & Bidan, G.) 98–205 (Elsevier, Amsterdam, 1999).
- Horowitz, G., Fichou, D., Peng, X., Xu, Z. & Garnier, F. *Solid State Comm.* **72**, 381–384 (1989).
- Garnier, F. *et al.* *J. Am. Chem. Soc.* **115**, 8716–8721 (1993).
- Dodabalapur, A., Torsi, L. & Katz, H. E. *Science* **268**, 270–271 (1995).
- Sirringhaus, H. *et al.* *Nature* **401**, 685–688 (1999).
- Bechgaard, K. & Jérôme, D. *Sci. Am.* **247**, 52–61 (1982).
- Lee, M., Schön, J. H., Kloc, C. & Batlogg, B. *Phys. Rev. Lett.* **86**, 862–865 (2001).

Daedalus

Spinning rubbish

A gas turbine, says Daedalus, is a collection of combustion chambers directing its flame onto a set of turbine blades. The shaft horsepower drives a compressor to feed the combustion process, and the exit gas forms a jet exhaust (in the jet engine). The turbine blades, directly in the path of the flame, have a severe duty, and can be internally cooled. A hollow blade with a divider along the middle can receive water through a hollow central shaft. In the centrifugal field the water will cling to the outside of the shaft bore, and even a small difference in density (caused, for example, by boiling) will circulate the water vigorously through the blade. Daedalus calculates that a fast-spinning turbine could drive the water in its blades supercritical at the blade tips.

Oxygenated, supercritical water is a powerful and spontaneous oxidizing agent. A turbine fed not with cooling water, but with an oxygenated fuel slurry, would therefore generate heat in its turbine blades by burning the fuel circulating in them. So DREADCO engineers are now building turbines whose combustion is distributed between the combustion chambers and the first few rows of blades. Indeed, Daedalus reasons that an internally heated blade could generate torque on its own account, making possible an entirely new sort of turbine.

The major application of turbines with water-loaded blades will be chemical, however. Many processes generate a weak slurry of some nasty organic chemical, and it is appealing to burn this in a turbine to a solution of CO₂ in water, and get useful power in exchange. The pharmaceutical industry, with innumerable waste streams on the way to a few useful products, would be an obvious customer. And the sewage business is dedicated to turning its customers' waste into a weak water slurry, and getting rid of this somehow. Burning it in turbines could generate vast amounts of electricity. Even the waste heat, now represented by the hot fizzy water rejected by the turbines, should be easy to discharge. Expensive cooling towers might not be needed.

Even better, many reactions of 'green chemistry' seek to replace messy industrial processes by neat supercritical ones from which the product can be elegantly recovered by simple processing. What a pity that supercritical dry cleaning, a splendid way of tidying up tatty or dirty assemblies, cannot be reduced to turbine processing.

David Jones

Obituary

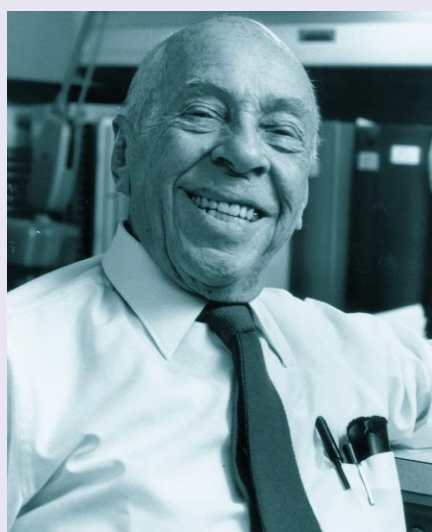
Ugo Fano (1912–2001)

Ugo Fano was a master at understanding how radiation interacts with matter. His work set the agenda for much of modern atomic physics, and had a broad sweep across fields as disparate as molecular biology and high-energy physics. His passing deprives us of a living link to the golden age of quantum mechanics, and to the “boys of the Via Panisperna” — Enrico Fermi’s remarkable school of physics.

A member of a wealthy Italian Jewish family, Fano was named after his grandfather, who served under Garibaldi in the war for Italian independence. His father, Gino, was a professor of geometry at the University of Turin, and his childhood was spent in a privileged social milieu that was shaken by the First World War, and shattered by the Second. A story well known to Fano’s colleagues recounts his first encounter at the age of 12 with Fermi: hiking along a mountain trail, Fano and his father were overtaken by a group of scientists from Rome. Pointing out a young man in the group, his father said: “That man is expected to go far in life. His name is Fermi.”

When Fano graduated from the University of Turin, he joined Fermi’s group in Rome at the age of 22. He soon made his most widely known contribution to science: an explanation of the asymmetric peaks seen in the spectra of radiation absorbed by atoms. This work was motivated by the spectral profiles of noble gases observed by Hans Beutler, but its results have found pervasive use. First published in *Nuovo Cimento* in 1935, the theory is best known in the extended form that Fano published in *Physical Review* (124, 1866–1878; 1961), which is still among that journal’s most-cited papers. Its simple ‘Fano profile’ formula, which predicts the shape of spectral lines, has been a workhorse of nuclear, atomic, molecular and condensed-matter physics. This work bears Fano’s distinctive signature, evocative of Fermi’s: complex physical phenomena are encapsulated in a few key parameters, which can be calculated from first principles.

The rise of fascism in Italy was at first regarded by Fano’s family as a matter of national embarrassment rather than as a direct personal threat. In 1936, Fano even joined Werner Heisenberg’s group in Leipzig, where, by his own account, he was treated as an Italian rather than as a Jew. But after his return to Rome in 1938, the



Outstanding interpreter of how radiation interacts with atoms and cells

Italian government accelerated its own descent into anti-Semitic madness. His family’s situation became untenable. After marrying Camilla Lattes in 1939, Fano went in exile to the United States, where he subsequently became a citizen.

His first paid job in America was in Milislav Demerec’s group at the Carnegie Institute’s Department of Genetics in Cold Spring Harbor. During his last days in Rome, Fano had been encouraged by Fermi to take up the study of biophysics from a fundamental perspective. The potent effects of X-rays on genetic materials suggested that radiation biology might identify the microscopic origin of the gene, a philosophy championed by Max Delbrück. Although trained as a theorist, Fano began studying the effects of X-rays on the eggs of the fruitfly *Drosophila melanogaster*, and, with Demerec, he performed the first isolation of *Escherichia coli* B.

Although Fano was soon drawn off to war work at the Aberdeen Proving Ground in Maryland, he derived a signal lesson from his forays into experimental biology. He realized that the effects of radiation on cells could not be encapsulated in a simple model of just the primary interaction. The energetic electrons produced by X-ray absorption initiate secondary reactions, and the ultimate effects of X-rays are determined by the dispersion of this energy in the material. Fano thus began a

systematic study of the ‘degradation’ of energetic radiation in matter. This required an understanding of radiative and electron interactions over a wide range of energies, and lasted for the rest of his scientific career.

After the war, Fano joined the staff of the National Bureau of Standards (NBS) in Washington, where he remained until 1966. He is thought to be the first ‘pure’ theoretical physicist hired by the NBS, and while there he made definitive contributions to the theory of spectral line shapes, the stopping power of charged particles, and the quantum mechanics of many-particle systems. His research group’s inclusion of women and African Americans would be noteworthy even by today’s standards, but in 1950s Washington it also had to contend with racial segregation of public facilities.

Fano was noted for his close and demanding interaction with experimental groups. In particular, he was a central figure in the NBS ‘resonance era’ of the early 1960s, during which novel, highly excited, atomic states were revealed in synchrotron radiation and electron scattering experiments. Fano’s analysis of the helium photoabsorption spectrum pointed to the existence of a correlated mode of two-electron excitation, which became a touchstone of his subsequent work.

Taking up an academic career at the University of Chicago in 1966, Fano established a school of atomic and molecular physics that addressed the blossoming experimental issues of that era, in particular those presented by the advent of laser spectroscopy. A number of now well-known effects are commonly identified by the adjective ‘Fano’, ranging from the production of spin-polarized electrons, to the excitation of deeply bound electrons in atomic collisions. Fano consistently applied his energies to the solution of great problems, and led his research group by personal inspiration.

Details of Fano’s accomplishments can be found in a forthcoming issue of *Physics Essays*, edited by Mitio Inokuti and Ravi Rau. He is survived by his wife, Camilla Lattes Fano; daughters Mary Fano Giacomoni and Virginia Fano Ghattas; four grandchildren, and a brother, Robert Fano. His many students and colleagues strive to emulate his approach to physics. **Charles W. Clark**
Charles W. Clark is in the Electron and Optical Physics Division, National Institute of Standards and Technology, Gaithersburg, Maryland 20899-8410, USA.
e-mail: charles.clark@nist.gov

Amyloid fibrils from muscle myoglobin

Even an ordinary globular protein can assume a rogue guise if conditions are right.

The sequence of amino acids in a natural protein determines the way in which it folds up into its unique biologically active 'native' conformation¹. But we show here that myoglobin, the archetypal globular protein², can convert into an alternative and radically different structure that closely resembles the amyloid and prion aggregates seen in pathological conditions such as Alzheimer's and Creutzfeldt–Jakob diseases^{3–6}. This most striking example of such a structural transition provides compelling evidence for the idea that amyloid represents a generic form of polypeptide conformation, but one that evolution has ensured is not normally formed in living systems^{7,8}.

Myoglobin (Fig. 1a) is a compact and highly soluble protein without any native-state properties to suggest that it has a predisposition to form amyloid fibrils. Whereas the latter are characteristically rich in β -sheets, native myoglobin lacks any elements of such structure and has most of its sequence arranged in well-defined α -helices. Moreover, all partially folded states of the protein characterized so far are significantly helical⁹, although, as with most proteins, there is evidence for the presence of more extended conformations in aggregates and precipitates^{9,10}.

In a screening process, in which pH, temperature and buffers were varied, we found that incubation of apomyoglobin (the protein without its haem group) in 50 mM sodium borate, pH 9.0 at 65 °C, conditions under which the native fold is substantially destabilized, resulted in the formation of large quantities of fibrillar structures (Fig. 1b). Solutions of these fibrils gave rise to a far-ultraviolet circular-dichroism spectrum containing a single minimum at 215 nm (Fig. 1c), which is diagnostic for the presence of β -structure. Moreover, an anisotropic X-ray diffraction pattern was obtained with characteristic 'cross- β ' reflections at 4.63 ± 0.08 and 10.11 ± 1.23 Å (Fig. 1d).

These results indicate that myoglobin fibrils contain β -strands that are oriented perpendicular to the main fibre axis⁵ and, together with the properties of the fibril-containing solutions in tinctorial assays using thioflavin T and Congo Red, indicate that these fibrils are indistinguishable in their core structure from disease-related amyloid fibrils⁵.

Our findings for myoglobin provide strong evidence that the sequences of polypeptides associated with amyloid or prion diseases need not be fundamentally distinct from those of other proteins by

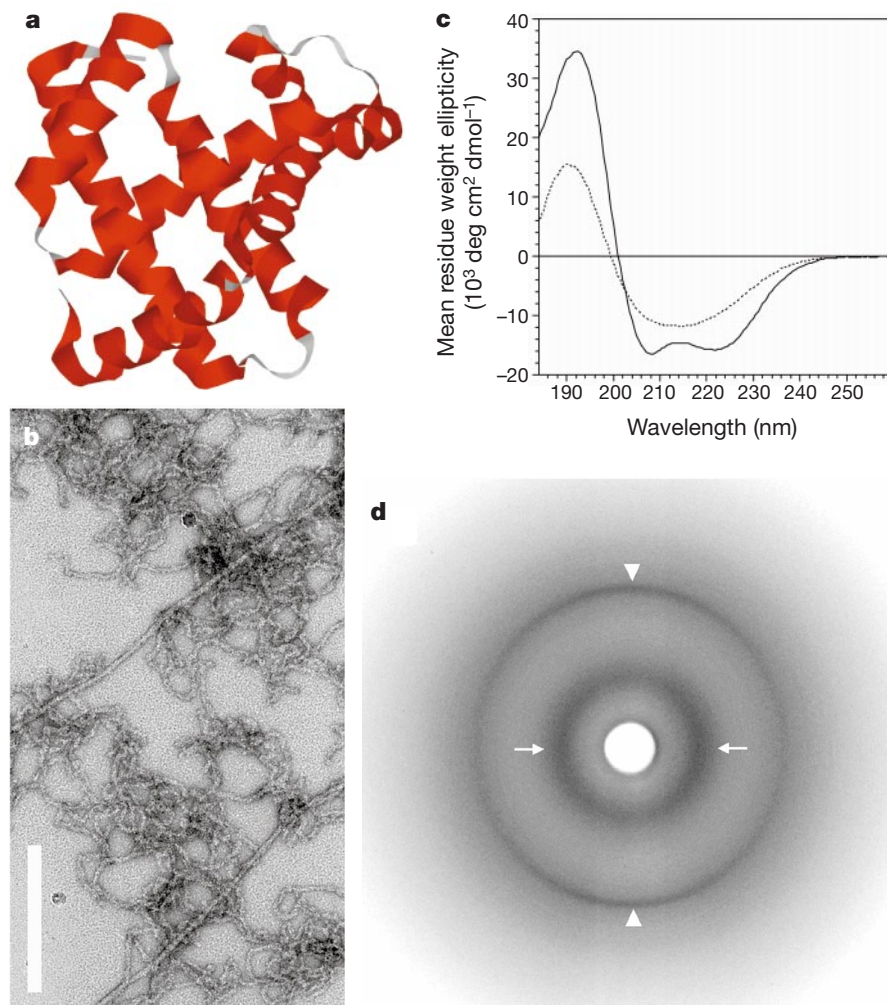


Figure 1 Structure of native and fibrillar myoglobin. **a**, Backbone representation of native myoglobin¹¹; α -helical residues are picked out in red. **b**, Electron micrograph of fibrils formed from horse skeletal-muscle myoglobin after haem extraction¹² and stained with uranyl acetate. Scale bar, 300 nm. **c**, Far-ultraviolet circular-dichroism spectra at 22 °C of freshly dissolved apomyoglobin (continuous line) and purified fibrils after incubation for 25 days (dotted line). **d**, X-ray diffraction pattern of fibres obtained using an 18-cm imaging plate detector (MarResearch) with a Rigaku RU200 rotating anode. Arrowheads (on the meridian) indicate the 4.63-Å reflection; arrows indicate the 10.11-Å reflection.

having any explicit coding for the cross- β structure⁷. But how can a polypeptide chain fold into two different, well-ordered states? In the native state of a globular protein such as myoglobin (Fig. 1a), side-chain interactions are crucial in defining the unique and characteristic main-chain fold¹. We suggest that, in contrast, the cross- β conformation is dominated by main-chain interactions that are common to different polypeptides; hence sequence effects are much less significant in determining the amyloid core structure. The latter is instead likely to be favoured by the inherent physical-chemical properties of the otherwise disordered polypeptide chains as they aggregate slowly under specific solution conditions in which

the native state is unstable.

We believe that the ability of natural protein sequences (which represent only a small fraction of all possible sequences) to fold efficiently into cooperative globular structures, together with other strategies such as employing molecular chaperones, is a very effective evolutionary adaptation to suppress amyloid formation *in vivo*. Conditions compromising such protective mechanisms, including ageing or mutational changes, may sometimes allow even a highly selected natural sequence to revert to its alternative conformation.

This proposal enables the basic principle of protein folding, namely that there is an unambiguous relationship between the

amino-acid sequence and the structure in the native state¹, to be reconciled with the evidence that even a protein such as myoglobin can adopt the fundamentally different but highly organized structure present in amyloid fibrils.

Marcus Fändrich, Matthew A. Fletcher, Christopher M. Dobson

Oxford Centre for Molecular Sciences,
New Chemistry Laboratory, University of Oxford,
South Parks Road, Oxford OX1 3QT, UK
e-mail: chris.dobson@chem.ox.ac.uk

1. Anfinsen, C. B. *Science* **181**, 223–230 (1973).

2. Kendrew, J. C. *et al.* *Nature* **185**, 422–427 (1960).
3. Tan, S. Y. & Pepys, M. B. *Histopathology* **25**, 403–414 (1994).
4. Rochet, J. C. & Lansbury, P. T. Jr *Curr. Opin. Struct. Biol.* **10**, 60–68 (2000).
5. Sunde, M. & Blake, C. C. F. *Q. Rev. Biophys.* **31**, 1–39 (1998).
6. Prusiner, S. B. *Proc. Natl Acad. Sci. USA* **95**, 13363–13383 (1998).
7. Dobson, C. M. *Trends Biochem. Sci.* **24**, 329–332 (1999).
8. Chiti, F. *et al.* *Proc. Natl Acad. Sci. USA* **96**, 3590–3594 (1999).
9. Eliezer, D., Yao, J., Dyson, H. J. & Wright, P. E. *Nature Struct. Biol.* **5**, 148–155 (1998).
10. Smeller, L., Rubens, P. & Heremans, K. *Biochemistry* **38**, 3816–3820 (1999).
11. Evans, S. V. & Brayer, G. D. *J. Biol. Chem.* **263**, 4263–4268 (1988).
12. Teale, F. W. J. *Biochim. Biophys. Acta* **35**, 543 (1959).

Pattern formation

Spiral cracks without twisting

A fascinating class of patterns, often encountered in nature as meandering cracks on rocks, dried-out fields and tectonic plates, is produced by the fracture of solids¹. Here we describe the observation and modelling of an unusual type of pattern consisting of spiral cracks in fragments of a thin layer of drying precipitate. We find that this symmetry-breaking cracking mode arises naturally not from twisting forces, but from a propagating stress front induced by the fold-up of the fragments.

Fractured surfaces and lines^{2–6} typically show a cellular and hierarchical pattern. Twisting forces produce a spiral fracture, like that often seen in a tibia bone broken in a skiing accident⁷. However, spiral cracks can also be created in other situations, as we show here by drying a fine aqueous suspension of precipitate. During drying, the suspension solidifies and later fragments into isolated parts (Fig. 1a).

Surprisingly, for very fine precipitates in a solidified layer of thickness between 0.2 and 0.5 mm, regular spiral as well as circular cracking pathways show up inside the fragments (Fig. 1b). Depending on the grain size, precipitate type and layer thickness, the size of the spirals varies widely from several hundred micrometres to a few millimetres. To the naked eye, they look like small dots, but their detail is revealed under a microscope (Fig. 1c). These spiral cracks do not occur in one particular material — we were able to generate them in three different precipitates, from nickel phosphate $\text{Ni}_3(\text{PO}_4)_2$, ferric ferrocyanide $\text{Fe}_4[\text{Fe}(\text{CN})_6]_3$ and ferric hydroxide $\text{Fe}(\text{OH})_3$.

Careful *in situ* observation suggested a mechanism for the formation of these cracks. The spirals and circle-shaped structures form only after the fragmentation process is over. Owing to the humidity gradient across the thickness, the fragments gradually fold up and detach from the substrate, generating large tensile stresses in the radial

direction, at and normal to the front of detachment. The extent of the attached area shrinks as the ring-shaped front advances inwards as a result of ongoing desiccation.

When the stress at the front exceeds the material strength, a crack is nucleated. As the nucleation is seldom symmetrical with respect to the boundaries, the crack tends to propagate along the front in only one direction, where more stresses can be released. By the time the crack growth completes a cycle, the front has already advanced, leading eventually to an inward spiral crack. As the stresses are concentrated at the layer–substrate interface, the spiral is confined there, with a typical penetration of 20–60% of the thickness.

The fact that the patterns are largely spiral suggests that crack propagation is favoured over nucleation, otherwise we would see more cylindrical concentric cracks. Although in a few instances we did observe this type of pattern, the majority are spiral in structure.

To test the proposed mechanism, we implemented it in a mesoscopic computer model⁸ that describes fracture on a frictional substrate. In this model, the grains in the

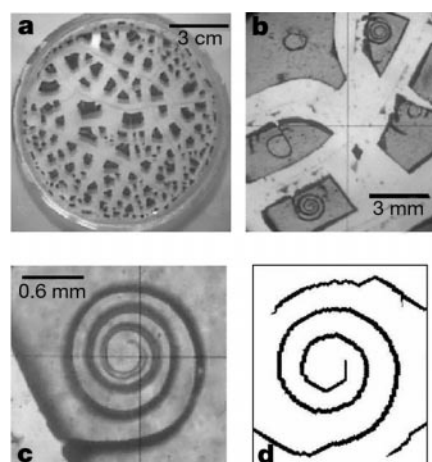


Figure 1 Spiral cracks revealed at small scales. **a**, Typical fragmentation pattern of nickel phosphate precipitate after desiccation in a Petri dish. **b**, Spiral and circular structures obtained inside the fragments. **c**, Close-up of a spiral crack. **d**, Computer-simulated spiral crack obtained using a mesoscopic spring-block model.

layer are represented by blocks on a triangular lattice, interconnected among neighbours by springs. The system is prestrained, and then relieved quasi-statically in a physical way. The relaxation is dictated by the competition between stick-and-slip and bond-breaking.

Focusing on the post-fragmentation process, we imposed a circular, inwardly propagating stress field to mimic the advancing detachment front. In a rather narrow parameter region, the spiral cracks were successfully reproduced (Fig. 1d). More tightly bound spirals can be obtained for smaller penetration depth, in agreement with experiment and prediction based on screening effects in the stress field. Other evidence has also been reported for the formation of similar spiral crack patterns under specific conditions⁹.

K.-t. Leung*, L. Józsa†, M. Ravasz‡, Z. Nédai§

*Institute of Physics, Academia Sinica, Taipei, Taiwan 11529, Republic of China
Departments of †Chemistry and ‡Physics,
Babes-Bolyai University, Cluj, RO-3400, Romania
§Department of Physics, University of Notre Dame,
Notre Dame, Indiana 46556, USA
e-mail: zneda@nd.edu

1. Lawn, B. *Fracture of Brittle Solids* 2nd edn (Cambridge Univ. Press, New York, 1993).
2. Skjeltorp, A. T. & Meakin, P. *Nature* **335**, 424–426 (1988).
3. Yuse, A. & Sano, M. *Nature* **362**, 329–331 (1993).
4. Groisman, A. & Kaplan, E. *Europhys. Lett.* **25**, 415–420 (1994).
5. Bai, T., Pollard, D. D. & Gao, H. *Nature* **403**, 753–756 (2000).
6. Shorlin, K. A., de Bruyn, J. R., Graham, M. & Morris, S. W. *Phys. Rev. E* **61**, 6950–6957 (2000).
7. Bostman, O. M. J. *Bone Joint Surg.* **68**, 462–466 (1986).
8. Leung, K.-t. & Nédai, Z. *Phys. Rev. Lett.* **85**, 662–665 (2000).
9. Xia, Z. C. & Hutchinson, J. W. *J. Mech. Phys. Solids* **48**, 1107–1131 (2000).

Polynesian origins

Slow boat to Melanesia?

The origin of the Polynesian islanders and of the Austronesian languages that they speak has been debated for more than 200 years. Diamond has presented the predominantly held modern viewpoint, described as the ‘express train to Polynesia’ model, which proposes that the ancestors of the Polynesians were early farmers who dispersed south from a homeland in South China/Taiwan, through Island Southeast Asia (replacing an indigenous ‘Australoid’ hunter-gatherer population), and then on east, out into the Pacific — all within the past 6,000 years¹. However, evidence is accumulating from several genetic markers that Polynesian lineages have a much deeper ancestry within tropical Island Southeast Asia than this hypothesis would suggest. The new evidence implies that the Polynesians originated not in China/Taiwan, but in eastern Indonesia, somewhere between Wallace’s line and the island of New Guinea.

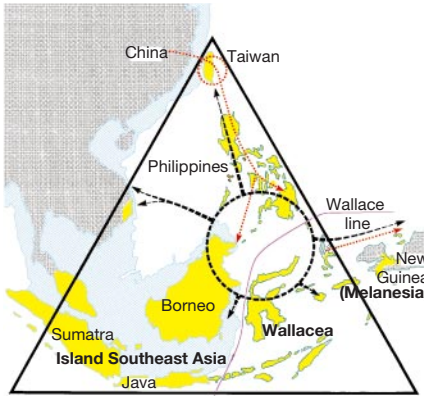


Figure 1 Map showing the two main alternative views of Austronesian origins. The oldest view represented by Meacham¹⁰ (triangle) and Solheim¹¹ (dashed black lines and circle) argues for an Island Southeast Asian homeland (before 5000 BC). Bellwood's⁵ and Diamond's¹ view of a recent rapid migration out of China (3000–4000 BC), spreading to replace all the older populations of Indonesia after 2000 BC is shown as a red dotted line. This is not supported by an ancient mitochondrial sequence haplotype, the 'Polynesian motif', found only to the east of Wallace's line (pink). Blue shading represents the continental shelf; yellow, areas where Austronesian languages are spoken.

Several genetic marker systems point to a primarily insular Southeast Asian ancestry for Polynesians. Much has been made of a hallmark maternal genetic marker known as the 'Polynesian motif'. This unique suite of four single-nucleotide polymorphisms in the control region of mitochondrial DNA identifies a subgroup within a widespread East Asian cluster of mitochondrial DNA lineages, haplogroup B, characterized by an intergenic 9-base-pair deletion^{2–4}. The Polynesian motif, so called because it reaches very high frequencies in Polynesian populations, is the main Oceanic variant of haplogroup B.

The Polynesian motif is also distributed throughout the lowland populations of coastal Melanesia and the biogeographic zone of Wallacea (Fig. 1)^{2,3}. More important, it is almost absent to the west of Wallace's line. It is not found in the Philippines, Taiwan or China — all key stations along the 'express train' route. Instead, in these regions its immediate ancestor is found with only three of the four polymorphisms, apparently breaking the train ride somewhere around Wallacea, where the final mutation in the motif, at nucleotide position 16,247, evidently occurred. This suggests that Wallacea, long believed to be an admixed buffer zone between Southeast Asia and New Guinea/Melanesia, might have harboured an ancient, indigenous population (of ultimately Asian origin) from which the Polynesian colonists emerged.

We looked into this possibility — that the final mutation did not occur en route in the express train, but earlier — by using the diversity accumulated by the motif to estimate its age using the molecular clock. The

motif dates back in Wallacea to roughly 17,000 years before present (95% credible region: 5,500–34,500 years)⁴. But archaeological evidence, mainly from using red-slipped pottery as a marker, argues for a tightly constrained arrival and departure of the express train from Wallacea around 2000 BC (ref. 5), suggesting that the motif originated before an express train carrying Taiwanese farmers could have arrived in Wallacea.

A report describing Y-chromosome variation in the region reaches a similar conclusion⁶, and earlier autosomal studies^{7,8} and physical anthropology⁹ also indicate ancient differentiation between mainland Asia, Taiwan, Island Southeast Asia and Melanesia. It is difficult to reconcile this evidence with the express train view — it seems to be more consistent with Austronesian origins within tropical Island Southeast Asia, as proposed earlier by other archaeologists^{10,11}. In particular, it points more to Polynesian origins between insular Southeast Asia and Melanesia, perhaps in the Pleistocene 'voyaging corridor'¹².

Diamond highlights evidence of linguistic diversity to support a Taiwanese origin for Austronesian languages, which identifies ten primary branches in Austronesian, nine of which are spoken only in Taiwan¹³. But the lack of an origin for the tenth branch in Taiwan weakens the case for a Taiwanese origin. The lack of equivalent deep-branch diversity in parts of Southeast Asia such as the Philippines may instead have resulted from the linguistic phenomenon of 'leveling'. If Taiwan had simply been an Austronesian backwater, as argued on the basis of archaeological evidence¹⁰, earlier levels of diversity might well have survived. Maybe the pre-Oceanic distribution of the Austronesian language family — its 'homeland' — was "... the broad triangular area formed by Taiwan, Sumatra, and Timor, where the reputedly oldest Malayo-Polynesian languages are found and where no other languages are spoken today"¹⁰.

Stephen J. Oppenheimer, Martin Richards*
Green College, Oxford OX2 6HG, UK
e-mail: stephen.oppenheimer@paediatrics.ox.ac.uk
*Department of Chemical and Biological Sciences, University of Huddersfield, Huddersfield HD1 3DH, UK

1. Diamond, J. M. *Nature* **403**, 709–710 (2000).
2. Melton, T. et al. *Am. J. Hum. Genet.* **57**, 403–414 (1995).
3. Redd, A. J. et al. *Mol. Biol. Evol.* **12**, 604–615 (1995).
4. Richards, M., Oppenheimer, S. J. & Sykes, B. *Am. J. Hum. Genet.* **63**, 1234–1236 (1998).
5. Bellwood, P. in *Prehistory of the Indo-Malaysian Archipelago* 2nd edn, 119 (University of Hawai'i Press, Honolulu, 1997).
6. Su, B. et al. *Proc. Natl Acad. Sci. USA* **97**, 8225–8228 (2000).
7. Serjeantson, S. W. & Hill, A. V. S. in *The Colonization of the Pacific: A Genetic Trail* (eds Hill, A. V. S. & Serjeantson, S. W.) 286 (Clarendon, Oxford, 1989).
8. Baer, A. S. *Asian Perspectives* **34**, 21–35 (1995).
9. Hanihara, T. *Anthropol. Sci.* **101**, 25–46 (1993).
10. Meacham, W. *Asian Perspectives* **26**, 89–106 (1984).
11. Solheim, W. G. II in *The History of Humanity* Vol. 1 (ed. De Laet, S. J.) 468–481 (Routledge, London, 1994).
12. Terrell, J. E. & Welsch, R. L. *Antiquity* **71**, 548–572 (1997).
13. Blust, R. *Symp. Ser. Inst. Linguist. Acad. Sinica* **1**, 31–94 (1999).

Diamond replies — Until about 4000 BC, there were no domestic animals, farming, Neolithic tools, or pottery anywhere in Island Southeast Asia and the Pacific, except for Neolithic agriculture in the New Guinea highlands. Around 4000 BC, all of these things appeared in Taiwan and then spread eastwards in archaeologically well-dated stages through the rest of Island Southeast Asia out into Polynesia. Apart from some peoples of New Guinea and some adjacent islands, all those in this vast realm now speak Austronesian languages.

Where did this Austronesian expansion ultimately originate? The archaeological, linguistic, zoological and botanical evidence points overwhelmingly to China, where all of the Austronesians' domestic animals, their pottery styles, their Neolithic tools, their irrigation and fishing technology, and many of their crops and artistic motifs originated. The spread south from China into Mainland Southeast Asia of all four of the other major Southeast Asian language families also correlates with archaeologically attested expansions.

Oppenheimer and Richards downplay this body of evidence, favouring instead an Austronesian origin in Island Southeast Asia itself, specifically in the triangle formed by Taiwan, Sumatra and Timor. But they overlook the fact that modern Austronesian peoples predominantly resemble Asian Mainland peoples in their genes, appearance and physical anthropology, whereas the original inhabitants of Island Southeast Asia (still attested by many relict populations today) resembled modern New Guineans and Aboriginal Australians.

Oppenheimer and Richards cite genetic evidence (for example, the Polynesian marker whose age they calculate), but if genetic evidence is to be useful in reconstructing human population movements, it must be based on many loci and integrated with other types of evidence; calculations of marker ages with extremely wide confidence limits must be viewed with caution. Nobody doubts that Melanesians and other original island peoples did make some minor contribution to the gene pool of the Austronesians as they expanded eastwards through the islands into Polynesia. The Polynesian marker, in combination with other genetic markers and other evidence, should eventually prove useful in illuminating where, when and how, among the hundreds of Austronesian peoples, the Polynesians themselves arose and received that minor contribution. But the marker cannot alter the main conclusions about the Austronesian expansion.

Jared Diamond
Physiology Department, University of California
School of Medicine, Los Angeles, California
90095-1751, USA
e-mail: jdiamond@mednet.ucla.edu

Nanotechnology

Thin solid films roll up into nanotubes

The rigorous size miniaturization of nanotechnology is continually generating new applications and new physical effects. We show here that nanotubes can be formed from thin solid films of almost any material at almost any position, once these films are released from their substrate. This exceptional design flexibility has useful implications, including for fluid transportation and capillarity on the nanometre scale, as well as offering the opportunity to extend fundamental investigations to a new diversity of materials, material systems and geometries.

There are two methods for making these solid-state nanotubes (Fig. 1a, b), involving a 'general' and a 'specialized' method. Both rely on the release of thin layers of the material from a substrate by a selective etching procedure. The position of the nanotubes is exactly determined by the etching time. Controlling the thickness of the layers and hence of the tube walls depends on the deposition method and can be as precise as a single atomic layer.

A wide range of possible materials can be used for this technique — calling only for appropriate deposition, a selective etchant and a certain amount of elasticity in the nanotube material. For example, single crystalline nanotubes can be created from semiconductor materials, which can be grown in perfect single-crystal quality.

A typical silicon–germanium nanotube prepared by a 'general' method (Fig. 1a) is shown in Fig. 1c. The nanotube is 12 μm long and 230 nm in diameter, which yields an aspect ratio of 60. The single-wall thickness (16 nm) is precisely defined by our deposition technique. Smaller diameters can be achieved with thinner layers: for example, we used a 6-nm SiGe layer to create a nanotube with a minimum diameter of only 50 nm (Fig. 1d), a feat helped by the remarkable elasticity of this material. The tube formed exactly at the edge of the back-etched buffer layer, demonstrating the unprecedented control our method offers over the nanotube's position on the substrate surface.

We were also able to create SiGe nanotubes by using a 'specialized' method (Fig. 1b). A SiGe-based nanotube, which has formed along the edge of the sample, is shown in Fig. 1e. This nanotube is 20 μm long and has a diameter of 530 nm. It adopts the slight bend of the sample edge and its position is again precisely defined by the etching time. Nanotubes of indium, gallium and arsenic have also been created by this method¹.

The variety of materials that can be

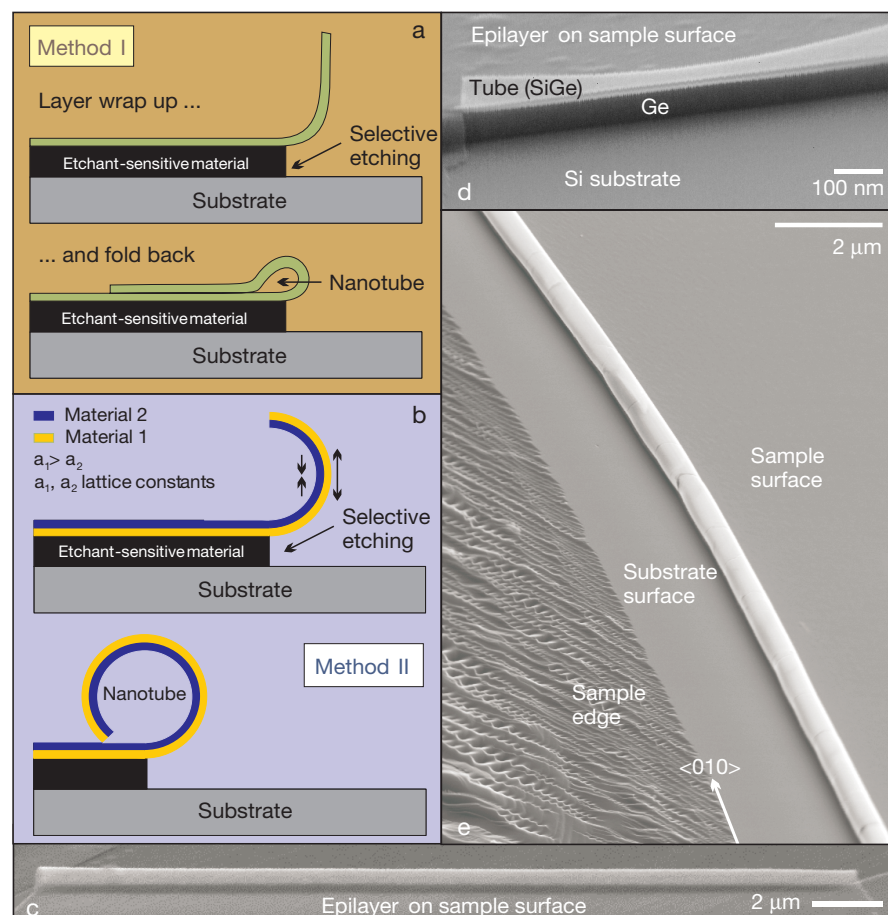


Figure 1 Formation of solid-state nanotubes. **a**, General method to create a nanotube (method I). An etchant-sensitive material is deposited on a substrate surface. On top of this sacrificial layer, a thin film (or a series of thin films) is deposited. After selective etching of the sacrificial layer, the thin top layer is wrapped up and folded back onto the sample surface, where it can bond to itself. At the position where the layer bends, a nanotube has formed. **b**, Specialized way to create a nanotube (method II). The layer sequence consists of an etchant-sensitive material, followed by a bilayer of two different materials (materials 1 and 2). Material 1 has a larger lattice constant than material 2, so $a_1 > a_2$. Once the bilayer is released by selective etching, each material tends to acquire its inherent lattice constant. The bilayer bends upwards, finally forming a nanotube after one complete revolution. Longer etching times result in multiple revolutions. **c**, **d**, Folded nanotubes fabricated according to method I. 'Epi' layer indicates epitaxial layer. **e**, SiGe-based nanotube formed along the edge of a sample (by method II).

engineered to produce solid-state nanotubes in this way, together with the precise control that can be exerted over their structure and position, make these nanotubes suitable candidates for use in fundamental investigations as well as in a range of applications. They could act as nanopipelines for fluid transportation and, as their diameter can be varied from nanometres to micrometres, they should be valuable in capillarity studies and in the design of syringe tips for nanoinjection, offering controlled generation of nanodroplets or nanobubbles in fluids or in fluid films on surfaces.

Nanotubes can act as nanodrillers, nanotweezers, microscopy tips, or as supporting rods in nanoobjects (if the wall is thick enough and the diameter small); if they have large diameters and thin walls, they can be used as nanocontainers or nanocapsules. Silicon-based compounds (SiC included) are particularly well suited for mechanical applications as this material is very hard and elastic.

Solid-state nanotubes could have electronic applications. Epitaxial growth (in which a layer of one material grows on a single crystal of another, so that the structure in the layer is the same as that in the substrate) can dope the grown layers in a precisely controlled manner over a concentration range of several powers of ten. Rolled-up nanotubes can serve as nanocables, or metal could be rolled into the nanotube or be made to fill it by capillary action. Possible roles in transistor design and in nanoelectronics have been proposed for carbon nanotubes^{2,3}, so the availability of alternative materials should offer an increased flexibility in their design and properties.

Oliver G. Schmidt, Karl Eberl

Max-Planck-Institut für Festkörperforschung,
Heisenbergstrasse 1, D-70569 Stuttgart,
Germany

e-mail: oschmidt@servix.mpi-stuttgart.mpg.de

1. Prinz, V. Y. et al. *Physica E* **6**, 828–831 (2000).

2. Tans, S. J., Verschueren, A. R. M. & Dekker, C. *Nature* **393**, 49–52 (1998).

3. Rueckes, T. et al. *Science* **289**, 94–97 (2000).

A measurement of the cosmological mass density from clustering in the 2dF Galaxy Redshift Survey

John A. Peacock¹, Shaun Cole², Peder Norberg², Carlton M. Baugh², Joss Bland-Hawthorn³, Terry Bridges³, Russell D. Cannon³, Matthew Colless⁴, Chris Collins⁵, Warrick Couch⁶, Gavin Dalton⁷, Kathryn Deeley⁶, Roberto De Propriis⁶, Simon P. Driver⁸, George Efstathiou⁹, Richard S. Ellis^{9,10}, Carlos S. Frenk², Karl Glazebrook¹¹, Carole Jackson⁴, Ofer Lahav⁹, Ian Lewis³, Stuart Lumsden¹², Steve Maddox¹³, Will J. Percival¹, Bruce A. Peterson⁴, Ian Price⁴, Will Sutherland^{1,7} & Keith Taylor^{3,10}

¹ Institute for Astronomy, University of Edinburgh, Royal Observatory, Blackford Hill, Edinburgh EH9 3HJ, UK

² Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK

³ Anglo-Australian Observatory, P.O. Box 296, Epping, New South Wales 2121, Australia

⁴ Research School of Astronomy & Astrophysics, The Australian National University, Weston Creek, ACT 2611, Australia

⁵ Astrophysics Research Institute, Liverpool John Moores University, Twelve Quays House, Birkenhead L14 1LD, UK

⁶ Department of Astrophysics, University of New South Wales, Sydney, New South Wales 2052, Australia

⁷ Department of Physics, University of Oxford, Keble Road, Oxford OX3RH, UK

⁸ School of Physics and Astronomy, University of St Andrews, North Haugh, St Andrews, Fife KY6 9SS, UK

⁹ Institute of Astronomy, University of Cambridge, Madingley Road, Cambridge CB3 0HA, UK

¹⁰ Department of Astronomy, Caltech, Pasadena, California 91125, USA

¹¹ Department of Physics & Astronomy, Johns Hopkins University, Baltimore, Maryland 21218-2686, USA

¹² Department of Physics, University of Leeds, Woodhouse Lane, Leeds LS2 9JT, UK

¹³ School of Physics & Astronomy, University of Nottingham, Nottingham NG7 2RD, UK

The large-scale structure in the distribution of galaxies is thought to arise from the gravitational instability of small fluctuations in the initial density field of the Universe. A key test of this hypothesis is that forming superclusters of galaxies should generate a systematic infall of other galaxies. This would be evident in the pattern of recessional velocities, causing an anisotropy in the inferred spatial clustering of galaxies. Here we report a precise measurement of this clustering, using the redshifts of more than 141,000 galaxies from the two-degree-field (2dF) galaxy redshift survey. We determine the parameter $\beta = \Omega^{0.6}/b = 0.43 \pm 0.07$, where Ω is the total mass-density parameter of the Universe and b is a measure of the 'bias' of the luminous galaxies in the survey. (Bias is the difference between the clustering of visible galaxies and of the total mass, most of which is dark.) Combined with the anisotropy of the cosmic microwave background, our results favour a low-density Universe with $\Omega \approx 0.3$.

Hubble showed in 1934 that the pattern of galaxies on the sky is non-random¹, and successive years have seen ever more ambitious attempts to map the distribution of visible matter on cosmological scales. In order to obtain a three-dimensional picture, redshift surveys use Hubble's law, $v = H_0 r$, to infer approximate radial distances to a set of galaxies. The first major surveys of this sort took place in the early 1980s (refs 2–5), and were limited to a few thousand redshifts, owing to the limited speed of single-object spectroscopy. In the 1990s, redshift surveys were extended to much larger volumes by a 'sparse sampling' strategy⁶. These studies^{7,8} established that the universe was close to uniform on scales above about $100 h^{-1} \text{ Mpc}$ (where $h \equiv H_0/100 \text{ km s}^{-1} \text{ Mpc}^{-1}$), but with a complex nonlinear supercluster network of walls, filaments and voids on smaller scales.

The origin of this large-scale structure is one of the key issues in cosmology. A plausible assumption is that structure grows by gravitational collapse of density fluctuations that are small at early times—but it is essential to test this idea. One important signature of gravitational instability is that collapsing structures should generate 'peculiar' velocities, δv , which distort the uniform Hubble expansion. We measure a redshift, z , which combines Hubble's law with the radial component of these peculiar velocities: $cz \approx H_0 r + \delta v \cdot \hat{r}$. The apparent density field seen in a redshift survey is thus not a true three-dimensional picture, but this can be turned to our advantage. The redshift-space distortions have a characteristic form, whose detection can both verify the general idea that structure forms by gravitational instability, and also measure the

density of the universe. Here we present measurements of this effect, based on a new large redshift survey.

The 2dF Galaxy Redshift Survey

New-generation large redshift surveys are made feasible by multiplexed fibre-optic spectroscopy, and the most advanced facility of this sort is the two-degree field, mounted at the prime focus of the Anglo–Australian Telescope⁹, which allows 400 spectra to be measured simultaneously. (For details of the 2df instrument, see <http://www.aao.gov.au/2df/>.) The 2dF Galaxy Redshift Survey (2dFGRS)¹⁰ was designed to use this instrument to measure the redshifts of 250,000 galaxies, to a blue magnitude limit of $b_1 = 19.45$ (corrected for extinction by dust in the Milky Way). (For details of the current status of the 2dFGRS, see <http://www.mso.anu.edu.au/2dFGRS/>.) The galaxies were selected from an updated version of the APM (automatic plate-measuring machine) catalogue¹¹, which is based on scans of photographic plates taken with the UK Schmidt telescope. Survey observations began in 1998, and should finish at the end of 2001. Redshifts have currently been obtained for 141,402 galaxies. The sky coverage of the 2dFGRS consists of strips in the northern and southern galactic poles ($75^\circ \times 7.5^\circ$ in the NGP; $75^\circ \times 15^\circ$ in the SGP), plus a number of outlying random fields. Coverage is now sufficiently extensive that near-complete thin slices through the galaxy distribution may be constructed, as shown in Fig. 1. This image illustrates well the median depth of the survey: approximately $z = 0.11$. Beyond this point, the survey is sensitive only to the more luminous galaxies, and the comoving density falls

rapidly. Nevertheless, the survey volume (including other regions not shown) is more than adequate for an accurate determination of the statistical properties of galaxy clustering.

Galaxy correlations in redshift space

The simplest statistical indicator of peculiar velocities in cosmological structure is the two-point correlation function, $\xi(\sigma, \pi)$. This measures the excess probability over random probability of finding a pair of galaxies with a transverse separation σ and a line-of-sight separation π . In an isotropic universe, this function should be independent of direction, but this is not true in redshift space. Transverse separations are true measures of distance, but apparent radial separations are distorted by peculiar velocities. This redshift-space anisotropy should cause two characteristic effects, operating respectively on small and large scales. On small scales, random orbital velocities within galaxy groups cause an apparent radial smearing, known as 'fingers of God'. Of greater interest is the large-scale effect; if cosmological structure forms by gravitational collapse, there should exist coherent infall velocities, and the effect of these is to cause an apparent flattening of structures along the line of sight. The general existence of redshift-space distortions was recognized in the first redshift surveys²⁻⁴, but the first comprehensive analysis of the phenomenon was performed by Kaiser¹², who showed that they could be used to measure the quantity

$$\beta \equiv \Omega^{0.6}/b$$

where Ω is the cosmological mass density parameter and b is the bias parameter that relates the relative density fluctuations of the galaxies and of the total mass:

$$\left. \frac{\delta\rho}{\rho} \right|_{\text{galaxies}} = b \left. \frac{\delta\rho}{\rho} \right|_{\text{mass}}$$

The presence of bias is an inevitable consequence of the nonlinear nature of galaxy formation, and the relation between mass and galaxy tracers is complex¹³⁻¹⁵. However, there are good theoretical reasons to expect that b can indeed be treated as a constant on large scales, where the density fluctuations are linear^{16,17}. Redshift-space

distortions have thus been seen as an important method for weighing the universe^{18,19}. To date, a number of papers have made significant detections of the Kaiser effect^{13,20,21}, but the 2dFGRS is the first survey that is large enough for the effect to be studied in detail.

In order to estimate $\xi(\sigma, \pi)$, we follow standard methods^{22,23} that compare the observed count of galaxy pairs with the count estimated using a random distribution that obeys the same selection effects in redshift and sky position. These selection effects are well defined, but complex: the survey is tessellated into a pattern of 'sectors' defined by the overlap of the 2° diameter survey tiles, whose positions are chosen adaptively with the aim of being able to place a fibre on more than 95% of the galaxies in the input catalogue. At the present intermediate stage of the survey, many tiles remain to be observed, and some regions of the survey presently contain redshifts for fewer than 50% of the galaxies. Furthermore, the spectroscopic success rate (redshifts per allocated fibre) is more than 95% in good conditions, but can fall to about 80% in marginal weather. We have implemented a number of independent algorithms for estimating the resulting survey selection effects, and are confident that we can measure the galaxy correlations robustly out to a separation of 25 h^{-1} Mpc. For example, the redshift distribution in sectors with low spectroscopic completeness is biased to low redshifts, but it makes no significant difference whether or not we correct for this, or indeed whether the low-completeness regions are simply excised. In addition to allowing for survey completeness, it is necessary to give higher weight to regions with a low sampling density, to achieve the optimum balance between cosmic variance and shot noise⁶. In practice, we have chosen to truncate the analysis at a maximum redshift of $z = 0.25$. Within this volume, the exact optimum weight per galaxy varies very nearly as the reciprocal of the number density, so that all volume elements receive approximately equal weight. The redshift-space correlation function for the 2dFGRS computed in this way is shown in Fig. 2. The correlation-function results display very clearly the two signatures of redshift-space distortions discussed above. The 'fingers of God' from small-scale random velocities are very clear, as indeed has been the case from the first redshift surveys³. However, this is the first time that the detailed

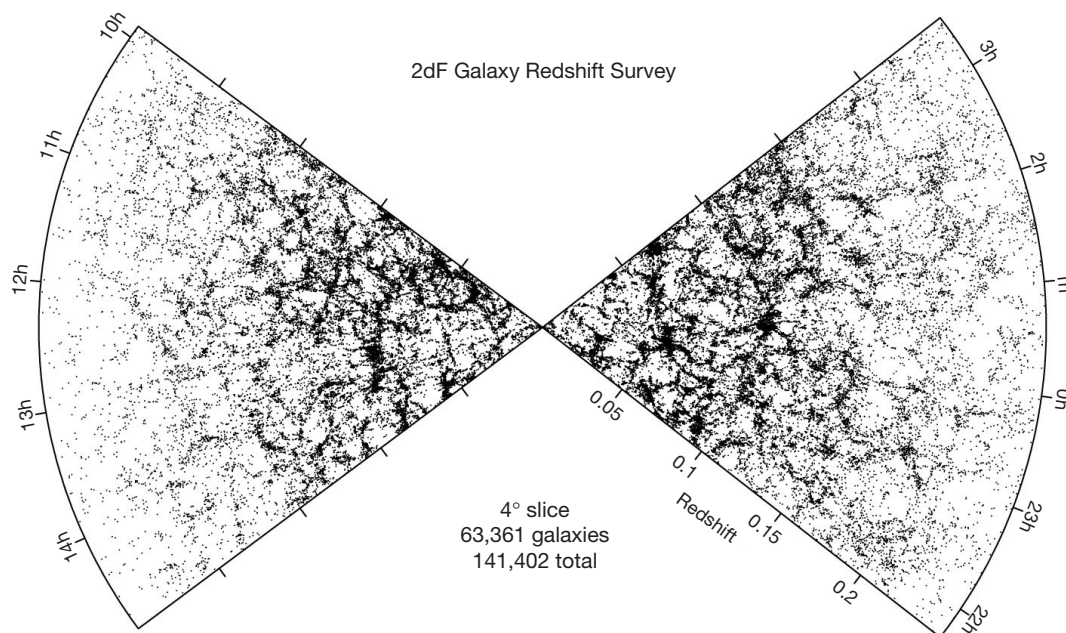


Figure 1 The distribution of galaxies in part of the two-degree-field galaxy redshift survey (2dFGRS), drawn from a total of 141,402 galaxies. The slices are 4° thick, centred at declination -2.5° in the Northern Galactic Pole (left) and -27.5° in the Southern Galactic Pole (right). Not all 2df fields within the slice have been observed at this stage; hence there are weak variations of the density of sampling as a function of right ascension. To

minimize such features, the slice thickness increases to 7.5° between right ascension 13.1h and 13.4h. This image reveals a wealth of detail, including linear supercluster features, often nearly perpendicular to the line of sight. The interesting question to settle statistically is whether such transverse features have been enhanced by infall velocities.

signature of large-scale flattening from coherent infall has been seen with a high signal-to-noise ratio.

Quantifying redshift-space distortions

The large-scale flattening of the correlation function may be quantified by measuring the quadrupole moment of $\xi(\sigma, \pi)$ as a function of radius r . A negative quadrupole moment implies flattening, whereas the finger-of-God distortion tends to yield a positive quadrupole moment. Figure 3 shows that the quadrupole-to-monopole ratio is positive on small scales, but that it falls with separation, becoming progressively more negative out to the largest separations at which it can be reliably measured. This arises partly because the underlying power spectrum is not a simple power-law function of scale, so that the peculiar velocities have a different effect at different radii. By integrating over the correlation function, it is possible to construct quantities in which this effect is eliminated. We shall not do this here, firstly because it seems desirable to keep the initial analysis as direct as possible. More importantly, finger-of-God smearing is a significant correction that will also cause the flattening to depend on radius. We therefore have to fit the data with a two-parameter model, described in the caption to Fig. 2. The parameters are β and a measure of the size of the random dispersion

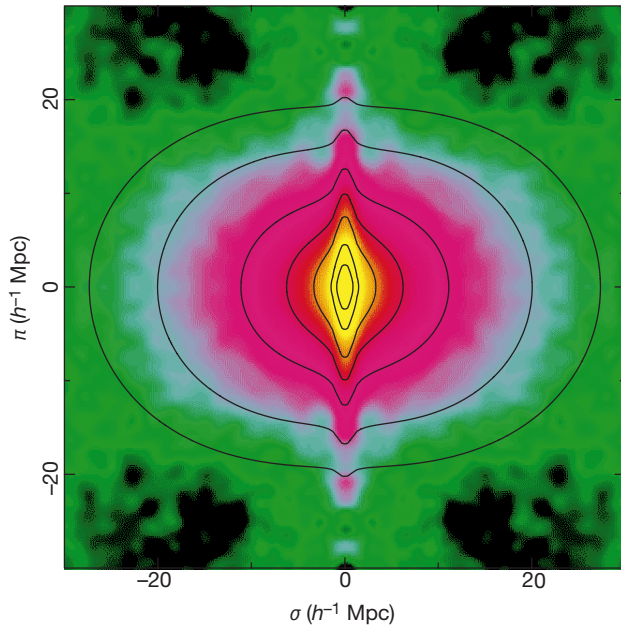


Figure 2 The redshift-space correlation function for the 2dFGRS, $\xi(\sigma, \pi)$, plotted as a function of transverse (σ) and radial (π) pair separation. The function was estimated by counting pairs in boxes of side $0.2 h^{-1}$ Mpc, and then smoothing with a gaussian of r.m.s. width $0.5 h^{-1}$ Mpc. To illustrate deviations from circular symmetry, the data from the first quadrant are repeated with reflection in both axes. This plot clearly displays redshift distortions, with finger-of-God elongations at small scales and the coherent Kaiser flattening at large radii. The overplotted contours show model predictions with flattening parameter $\beta = \Omega^{0.6}/b = 0.4$ and a pairwise dispersion of $\sigma_p = 400 \text{ km s}^{-1}$. Contours are plotted at $\xi = 10, 5, 2, 1, 0.5, 0.2$ and 0.1 .

The model predictions assume that the redshift-space power spectrum (P_s) may be expressed as a product of the linear Kaiser distortion and a radial convolution³⁷: $P_s(\mathbf{k}) = P_r(k)(1 + \beta\mu^2)^2(1 + k^2\sigma_p^2\mu^2/2H_0^2)^{-1}$, where $\mu = \hat{\mathbf{k}} \cdot \hat{\mathbf{r}}$, and σ_p is the r.m.s. pairwise dispersion of the random component of the galaxy velocity field. This model gives a very accurate fit to exact nonlinear simulations³³. For the real-space power spectrum, $P_r(k)$, we take the estimate obtained by deprojecting the angular clustering in the APM survey^{11,39}. This agrees very well with estimates that can be made directly from the 2dFGRS, as will be discussed elsewhere. We use this model only to estimate the scale dependence of the quadrupole-to-monopole ratio (although Fig. 2 shows that it does match the full $\xi(\sigma, \pi)$ data very well).

in the relative velocities of galaxies, σ_p . In practice, σ_p plays the role of an empirical fitting parameter to describe the scale on which the distortions approach the linear-theory predictions. It therefore also incorporates other possible effects, such as a scale dependence of bias.

The results for the quadrupole-to-monopole ratio are shown in Fig. 3, which shows the average of the estimates for the NGP and SGP slices. The difference between the NGP and SGP allows an estimate of the errors to be made: these slices are independent samples for the present analysis of clustering on relatively small scales. For model fitting, it is necessary to know the correlation between the values at different r . A simple way of addressing this is to determine the effective number of degrees of freedom from the value of χ^2 for the best-fitting model. A more sophisticated approach is to generate realizations of $\xi(\sigma, \pi)$, and construct the required covariance matrix directly. One way of achieving this is to analyse large numbers of mock surveys drawn from numerical simulations²⁴. A more convenient method is to generate direct realizations of the redshift-space power spectrum, using gaussian fluctuations on large scales, but allowing for enhanced variance in power on nonlinear scales^{25–27}. In practice, the likelihood contours resulting from this approach agree well with those from the simple approach, and we are confident that the resulting errors on β are realistic. These contours are shown in Fig. 4, and show that there is a degree of correlation between the preferred values of β and σ_p , as expected. For our purposes, σ_p is an uninteresting parameter, so we marginalize over it to obtain the following estimate of β and its root mean square (r.m.s.) uncertainty:

$$\beta = 0.43 \pm 0.07$$

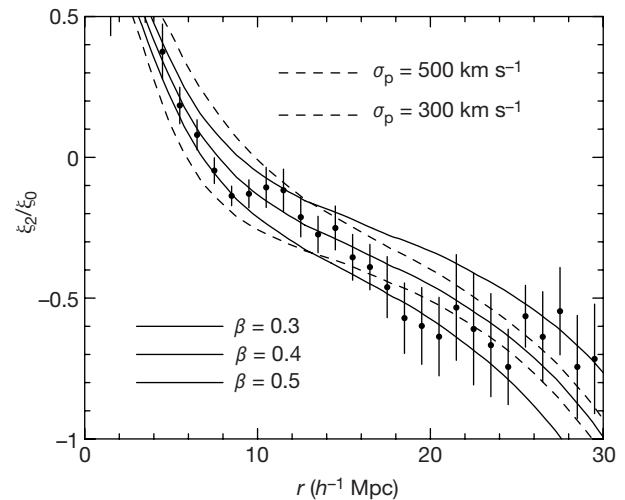


Figure 3 The flattening of the redshift-space correlation function is quantified by the quadrupole-to-monopole ratio, ξ_2/ξ_0 . This quantity is positive where fingers-of-God distortion dominates, and is negative where coherent infall dominates. The solid lines show model predictions for $\beta = 0.3, 0.4$ and 0.5 , with a pairwise velocity dispersion of $\sigma_p = 400 \text{ km s}^{-1}$ (solid lines), plus $\beta = 0.4$ with $\sigma_p = 300$ and 500 km s^{-1} (dashed lines). The ξ_2/ξ_0 ratio becomes more negative as β increases and as σ_p decreases. At large radii, the effects of fingers-of-God become relatively small, and values of $\beta \approx 0.4$ are clearly appropriate.

The multipole moments of the correlation function are defined as $\xi_\ell(r) = (2\ell + 1)/2\int_{-1}^1 \xi(\sigma = r \sin \theta, \pi = r \cos \theta) P_\ell(\cos \theta) d \cos \theta$. In linear theory, the quadrupole-to-monopole ratio is given⁴⁰ by $\xi_2/\xi_0 = f(n)(4\beta/3 + 4\beta^2/7)/(1 + 2\beta/3 + \beta^2/5)$. Here $f(n) = (3 + n)/n$, where n is the power-spectrum index of the density fluctuations: $\xi \propto r^{-(3+n)}$. In practice, nonlinear effects mean that this ratio is a function of scale. We model this by using the real-space correlation function estimated from the APM survey^{11,39}, plus the model for nonlinear finger-of-God smearing given in the caption to Fig. 2.

This result is the first precise determination of β from redshift-space distortions. The best previous studies^{13,20,21} have only achieved a detection of the effect at the 3σ level. Before discussing the implications of our result, we should therefore consider some possible small systematic corrections that have been unimportant in earlier work. First, the Kaiser analysis applies only in the small-angle approximation, and in principle corrections might be needed for wide-angle surveys such as ours²⁸. However, with our weighting scheme, the mean angular separation of pairs with spatial separations less than $30 h^{-1} \text{ Mpc}$ is only 2.5° , so this is not a concern. There is potentially a significant correction for luminosity effects. The optimal weighting means that our mean luminosity is high: it is approximately $M_{b_i} = -20.3$, or 1.9 times the characteristic luminosity, L^* , of the overall galaxy population²⁹. Several studies³⁰ have suggested that the strength of galaxy clustering increases with luminosity, with an effective bias that can be fitted by $b/b^* = 0.7 + 0.3(L/L^*)$. This effect has been controversial³¹, but the 2dFGRS data set favours a similar luminosity dependence, as will be described elsewhere. We therefore expect that β for L^* galaxies will exceed our directly measured figure. Applying a correction using the given formula for $b(L)$, we deduce

$$\beta(L = L^*) = 0.54 \pm 0.09$$

Finally, the 2dFGRS has a median redshift of 0.11. With weighting, the mean redshift in the present analysis is $\bar{z} = 0.17$, and our measurement should be interpreted as β at that redshift. The extrapolation to $z = 0$ is model-dependent, but probably does not introduce a significant change³².

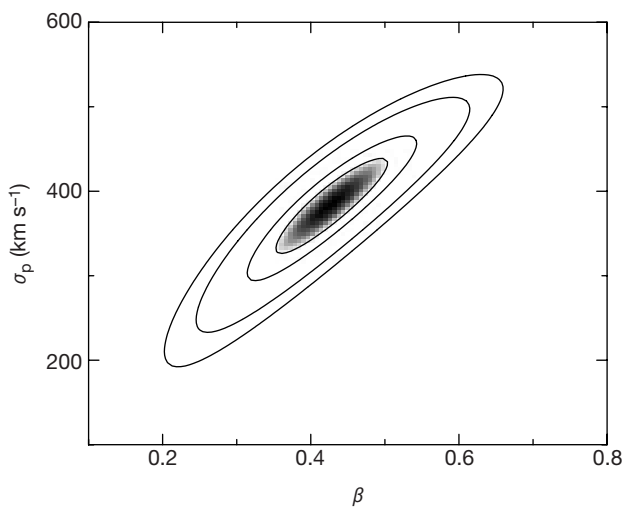


Figure 4 Likelihood contours for β and the fingers-of-God smearing parameter σ_p , based on the data in Fig. 3 (considering $8 h^{-1} \text{ Mpc} < r < 25 h^{-1} \text{ Mpc}$). These are plotted at the usual positions for one-parameter confidence of 68% (shaded region), and two-parameter confidence of 68%, 95% and 99% (that is, $\Delta\chi^2 = 1, 2.3, 6.0$ and 9.2). The maximum-likelihood solution is $\beta = 0.43$ and $\sigma_p = 385 \text{ km s}^{-1}$. The value for the large-scale pairwise dispersion is in reasonable agreement with previously suggested values⁴¹; however, for the present analysis σ_p is an uninteresting parameter. If we marginalize over σ_p (that is, integrate over σ_p , treating the likelihood as a probability distribution), the final estimate of β and its r.m.s. uncertainty is $\beta = 0.43 \pm 0.07$. We believe that this result is robust, in the sense that systematic errors in the modelling are smaller than the random errors. We have tried assuming that the power spectrum for $k < 0.1 h \text{ Mpc}^{-1}$ has the shape of a $\Omega = 0.3$ CDM model, rather than the APM measurement; this has a very small effect. A more serious issue is whether the pairwise velocity dispersion of galaxies may depend strongly on separation, as is found for mass particles in numerical simulations⁴². Assuming that the pairwise velocity dispersion σ_p rises to twice its large-scale value below $1 h^{-1} \text{ Mpc}$ reduces the best-fit β by 0.04. This correction is small because our analysis excludes the nonlinear data at $r < 8 h^{-1} \text{ Mpc}$.

Consistency with microwave-background anisotropies

We have verified in some detail that the pattern of redshift-space distortions associated with the gravitationally driven growth of clustering exists as predicted. Although gravitational instability is well established as the standard model for the formation of large-scale structure, it is important to have verified such a characteristic feature of the theory. Extracting the full cosmological implications of our measurement of $\Omega^{0.6}/b$ requires us to know the bias parameter in order to determine Ω . For example, our measurement implies $\Omega = 0.36 \pm 0.10$ if L^* galaxies are unbiased, but it is difficult to justify such an assumption. In principle, the details of the clustering pattern in the nonlinear regime allow the Ω - b degeneracy to be broken, yielding a direct determination of the degree of bias³³. We expect to pursue this approach using the 2dFGRS. For the present, however, it is interesting to use an independent approach. Observations of anisotropies in the cosmic microwave background (CMB) can in principle measure almost all the cosmological parameters, and current small-scale anisotropy results are starting to tighten the constraints. In a recent analysis³⁴, best-fitting values for the densities in collisionless matter, baryons and vacuum (subscripts c, b and v, respectively) have been obtained: $\Omega_c + \Omega_b + \Omega_v = 1.11 \pm 0.07$, $\Omega_c h^2 = 0.14 \pm 0.06$, $\Omega_b h^2 = 0.032 \pm 0.005$, together with a power-spectrum index $n = 1.01 \pm 0.09$. Our result for β gives an independent test of this picture, as follows.

The only parameter left undetermined by the CMB data is the Hubble constant, h . Recent work^{35,36} indicates that this is now determined to an r.m.s. accuracy of 10%, and we adopt a central value of $h = 0.70$. This completes the cosmological model, requiring a total matter density parameter $\Omega = \Omega_c + \Omega_b = 0.35 \pm 0.14$. It is then possible to use the parameter limits from the CMB to predict a conservative range for the mass power spectrum at $z = 0$, which is shown in Fig. 5. In this plot, the mass power spectrum appears to be in good agreement with the clustering observed in the APM survey. For each model allowed by the CMB, we can predict both b (from the ratio of galaxy and mass spectra) and also β (because a given CMB model specifies Ω). In practice, we determine

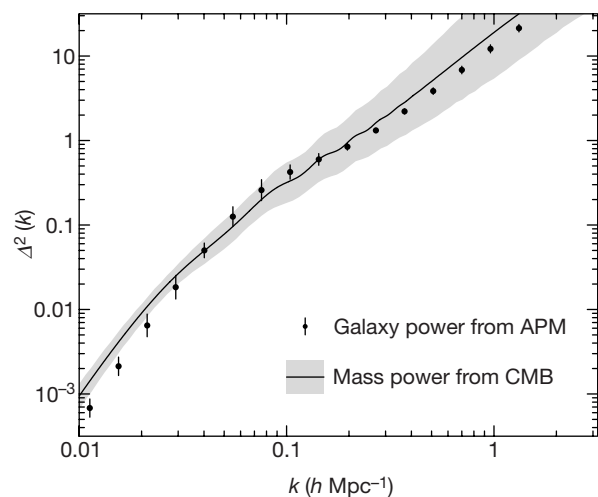


Figure 5 The dimensionless matter power spectrum at zero redshift, $\Delta^2(k)$, as predicted from the allowed range of models that fit the microwave-background anisotropy data, plus the assumption that $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1} \pm 10\%$. The solid line shows the best-fit model³⁴ (power-spectrum index $n = 1.01$, and density parameters in baryons, CDM, and vacuum of, respectively, 0.065, 0.285 and 0.760). The effects of nonlinear evolution have been included⁴³. The shaded points are the real-space power spectrum measured for APM galaxies. The clear conclusion is that APM galaxies are consistent with being essentially unbiased tracers of the mass on large scales. As the CMB data also constrain the range of Ω , this allows β to be predicted.

b by determining the mean ratio of power spectra over the range $0.02 < k < 0.1 h \text{Mpc}^{-1}$, where the APM measurement is robust and where scale-dependent bias and nonlinearities should be unimportant. Considering the allowed range of models, we then obtain the prediction $\beta_{\text{CMB+APM}} = 0.57 \pm 0.17$. A flux-limited survey such as the APM will have a mean luminosity close to L^* , so the appropriate comparison is with the 2dFGRS corrected figure of $\beta = 0.54 \pm 0.09$ for L^* galaxies. These numbers are in very close agreement.

This analysis of galaxy clustering in the 2dFGRS thus gives strong support to the simplest picture of cosmological structure formation, in which the primary mechanism is gravitational instability in a sea of collisionless dark matter. We have shown that the fluctuations seen in the CMB (which measure structure at a redshift $z \approx 1100$) can be extrapolated to the present to predict the peculiar velocities that distort redshift-space clustering. The agreement between this extrapolation and direct observations from the 2dFGRS is a highly non-trivial confirmation of the basic model. The precision of data in both areas should improve rapidly, and the use of β as a meeting ground between studies of the CMB and large-scale structure will undoubtedly lead to more demanding tests of the theory in years to come. For the present, we can say that there is complete consistency between clustering in the 2dFGRS and the emerging 'standard model' of cosmology: a spatially flat, vacuum-dominated universe with density parameter $\Omega \approx 0.3$. $\square$

Received 20 December 2000; accepted 30 January 2001.

1. Hubble, E. P. The distribution of extra-galactic nebulae. *Astrophys. J.* **79**, 8–76 (1934).
2. Kirshner, R. P., Oemler, A., Schechter, P. L. & Shectman, S. A. A million cubic megaparsec void in Bootes. *Astrophys. J.* **248**, L57–L60 (1981).
3. Davis, M. & Peebles, P. J. E. A survey of galaxy redshifts. V—The two-point position and velocity correlations. *Astrophys. J.* **267**, 465–482 (1983).
4. Bean, A. J., Ellis, R. S., Shanks, T., Efstathiou, G. & Peterson, B. A. A complete galaxy redshift sample. I—The peculiar velocities between galaxy pairs and the mean mass density of the Universe. *Mon. Not. R. Astron. Soc.* **205**, 605–624 (1983).
5. de Lapparent, V., Geller, M. J. & Huchra, J. P. A slice of the universe. *Astrophys. J.* **302**, L1–L5 (1986).
6. Kaiser, N. A sparse-sampling strategy for the estimation of large-scale clustering from redshift surveys. *Mon. Not. R. Astron. Soc.* **219**, 785–790 (1986).
7. Saunders, W. et al. The density field of the local universe. *Nature* **349**, 32–38 (1991).
8. Shectman, S. A. et al. The Las Campanas redshift survey. *Astrophys. J.* **470**, 172–188 (1996).
9. Taylor, K., Cannon, R. D. & Watson, F. G. Anglo-Australian Telescope's 2dF facility. *Proc. SPIE* **2871**, 145–149 (1999).
10. Colless, M. First results from the 2dF galaxy redshift survey. *Phil. Trans. R. Soc. Lond. A* **357**, 105–116 (1999).
11. Maddox, S., Efstathiou, G. & Sutherland, W. J. The APM Galaxy Survey—III. An analysis of systematic errors in the angular correlation function and cosmological implications. *Mon. Not. R. Astron. Soc.* **283**, 1227–1263 (1996).
12. Kaiser, N. Clustering in real space and in redshift space. *Mon. Not. R. Astron. Soc.* **227**, 1–21 (1987).
13. Hamilton, A. J. S., Tegmark, M. & Padmanabhan, N. *Mon. Not. R. Astron. Soc.* **317**, L23–L27 (2000).
14. Pen, U.-L. Reconstructing nonlinear stochastic bias from velocity space distortions. *Astrophys. J.* **504**, 601–606 (1998).
15. Dekel, A. & Lahav, O. Stochastic nonlinear galaxy biasing. *Astrophys. J.* **520**, 24–34 (1999).
16. Benson, A. J., Cole, S., Frenk, C. S., Baugh, C. M. & Lacey, C. G. The nature of galaxy bias and clustering. *Mon. Not. R. Astron. Soc.* **311**, 793–808 (2000).
17. Kauffmann, G., Nusser, A. & Steinmetz, M. Galaxy formation and large-scale bias. *Mon. Not. R. Astron. Soc.* **286**, 795–811 (1997).

18. Strauss, M. A. & Willick, J. A. The density and peculiar velocity fields of nearby galaxies. *Phys. Rep.* **261**, 271–431 (1995).
19. Hamilton, A. J. S. in *The Evolving Universe* Vol. 185, 231 (Kluwer Astrophysics and Space Science Library, Kluwer, Dordrecht, 1998); also preprint astro-ph/9708102 at (xxx.lanl.gov) (1997).
20. Taylor, A. N., Ballinger, W. E., Heavens, A. F. & Tadros, H. Application of data compression methods to the redshift-space distortions of the PSCz galaxy catalogue. *Mon. Not. R. Astron. Soc.* (in the press); also preprint astro-ph/0007048 at (xxx.lanl.gov) (2000).
21. Outram, P. J., Hoyle, F. & Shanks, T. The Durham/UKST galaxy redshift survey—VII. Redshift-space distortions in the power spectrum. *Mon. Not. R. Astron. Soc.* (in the press); also preprint astro-ph/0009387 at (xxx.lanl.gov) (2000).
22. Hamilton, A. J. S. Toward better ways to measure the galaxy correlation function. *Astrophys. J.* **417**, 19–35 (1993).
23. Landy, S. D. & Szalay, A. S. Bias and variance of angular correlation functions. *Astrophys. J.* **412**, 64–71 (1993).
24. Cole, S., Hatton, S., Weinberg, D. H. & Frenk, C. S. Mock 2dF and SDSS galaxy redshift surveys. *Mon. Not. R. Astron. Soc.* **300**, 945–966 (1998).
25. Feldman, H. A., Kaiser, N. & Peacock, J. A. Power-spectrum analysis of three-dimensional redshift surveys. *Astrophys. J.* **426**, 23–37 (1994).
26. Meiksin, A. A. & White, M. The growth of correlations in the matter power spectrum. *Mon. Not. R. Astron. Soc.* **308**, 1179–1184 (1999).
27. Scoccimarro, R., Zaldarriaga, M. & Hui, L. Power spectrum correlations induced by nonlinear clustering. *Astrophys. J.* **527**, 1–15 (1999).
28. Szalay, A. S., Matsubara, T. & Landy, S. D. Redshift-space distortions of the correlation function in wide-angle galaxy surveys. *Astrophys. J.* **498**, L1–L4 (1998).
29. Folkes, S. J. et al. The 2dF galaxy redshift survey: spectral types and luminosity functions. *Mon. Not. R. Astron. Soc.* **308**, 459–472 (1999).
30. Benoist, C., Maurogordato, S., da Costa, L. N., Cappi, A. & Schaeffer, R. Biasing in the galaxy distribution. *Astrophys. J.* **472**, 452–459 (1996).
31. Loveday, J., Maddox, S. J., Efstathiou, G. & Peterson, B. A. The Stromlo-APM redshift survey. 2: Variation of galaxy clustering with morphology and luminosity. *Astrophys. J.* **442**, 457–468 (1995).
32. Carlberg, R. G. et al. Galaxy clustering evolution in the CNOC2 high-luminosity sample. *Astrophys. J.* **542**, 57–67 (2000).
33. Verde, L., Heavens, A. F., Matarrese, S. & Moscardini, L. Large-scale bias in the Universe—II. Redshift-space bispectrum. *Mon. Not. R. Astron. Soc.* **300**, 747–756 (1998).
34. Jaffe, A. et al. Cosmology from Maxima-1, Boomerang and COBE/DMR CMB Observations; preprint astro-ph/0007333 at (xxx.lanl.gov) (2000).
35. Mould, J. R. et al. The Hubble Space Telescope key project on the extragalactic distance scale. XXVIII. Combining the constraints on the Hubble constant. *Astrophys. J.* **529**, 786–794 (2000).
36. Freedman, W. L. et al. Final results from the Hubble Space Telescope key project to measure the Hubble constant; preprint astro-ph/0012376 at (xxx.lanl.gov) (2000).
37. Ballinger, W. E., Peacock, J. A. & Heavens, A. F. Measuring the cosmological constant with redshift surveys. *Mon. Not. R. Astron. Soc.* **282**, 877–888 (1996).
38. Hatton, S. & Cole, S. Modelling the redshift-space distortion of galaxy clustering. *Mon. Not. R. Astron. Soc.* **296**, 10–20 (1998).
39. Baugh, C. M. & Efstathiou, G. The three-dimensional power spectrum measured from the APM galaxy survey—2. Use of the two-dimensional power spectrum. *Mon. Not. R. Astron. Soc.* **267**, 323–332 (1994).
40. Hamilton, A. J. S. Measuring omega and the real correlation function from the redshift correlation function. *Astrophys. J.* **385**, L5–L8 (1992).
41. Jing, Y. P., Mo, H. J. & Börner, G. Spatial correlation function and pairwise velocity dispersion of galaxies: Cold dark matter models versus the Las Campanas survey. *Astrophys. J.* **494**, 1–12 (1998).
42. Jenkins, A. R. et al. Evolution of structure in cold dark matter universes. *Astrophys. J.* **499**, 20–40 (1998).
43. Peacock, J. A. & Dodds, S. J. Non-linear evolution of cosmological power spectra. *Mon. Not. R. Astron. Soc.* **280**, L19–L26 (1996).

Acknowledgements

The 2dF Galaxy Redshift Survey was made possible through the dedicated efforts of the staff of the Anglo-Australian Observatory, both in creating the two-degree-field instrument and in supporting it on the telescope.

Correspondence and requests for materials should be addressed to J.A.P. (e-mail: jap@roe.ac.uk).

Defining brain wiring patterns and mechanisms through gene trapping in mice

Philip A. Leighton^{*,†}, Kevin J. Mitchell^{*,†,‡}, Lisa V. Goodrich^{*,†}, Xiaowei Lu[†], Kathy Pinson[‡], Paul Scherz[‡], William C. Skarnes[‡] & Marc Tessier-Lavigne[†]

[†] Howard Hughes Medical Institute and Departments of Anatomy and of Biochemistry and Biophysics, University of California, San Francisco, California 94143-0452, USA

[‡] Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA

* These authors contributed equally to this work

The search to understand the mechanisms regulating brain wiring has relied on biochemical purification approaches in vertebrates and genetic approaches in invertebrates to identify molecular cues and receptors for axon guidance. Here we describe a phenotype-based gene-trap screen in mice designed for the large-scale identification of genes controlling the formation of the trillions of connections in the mammalian brain. The method incorporates an axonal marker, which helps to identify cell-autonomous mechanisms in axon guidance, and has generated a resource of mouse lines with striking patterns of axonal labelling, which facilitates analysis of the normal wiring diagram of the brain. Studies of two of these mouse lines have identified an *in vivo* guidance function for a vertebrate transmembrane semaphorin, *Sema6A*, and have helped re-evaluate that of the Eph receptor *EphA4*.

Biochemical purification strategies in vertebrates and genetic screens in invertebrates have converged on several conserved families of molecular cues and receptors for axon guidance¹. In addition to these initial discoveries, a thorough understanding of human brain wiring mechanisms is likely to require genetic screens in mammals. Mammals exhibit a greater diversity of axon guidance molecules than invertebrates; for example, the neuropilin family of semaphorin receptors and the chemoattractant hepatocyte growth factor do not have invertebrate orthologues^{2,3}. Furthermore, biochemical strategies are not well suited to a systematic dissection of brain wiring, as assays must be tailored to each guidance event and require sufficient expression of the factor(s) being purified. By contrast, genetic approaches are not limited by expression levels and can identify molecules important for guidance of both small and large populations of axons.

Conventional genetic screens for brain wiring defects in mice are, however, daunting. Screening random chemically induced mutations can be inefficient, as only a small fraction of mutants are likely to present defects in a process as specific as axon guidance^{4–6}; furthermore, many of these defects will arise only indirectly, for example as a result of mutations in genes regulating the development of tissues through which axons navigate⁷. This problem presents a serious impediment in mice because of cost and the time-consuming task of identifying the mutated gene by positional cloning. Second, many mutants will display subtle wiring defects affecting small numbers of axons, leaving most of the nervous system unaffected (for example, see refs 8, 9). Finding such specific defects is often challenging or impossible, as many axon tracts lack specific markers to visualize them against the background of a mostly normal pattern of projections. Indeed, finding specific wiring defects is tantamount to finding the proverbial needle in a haystack. Finally, little is known about the normal wiring pattern of the brain and its development, often making it difficult to detect or assess wiring defects.

Here we report a modified 'gene-trapping' method, the PLAP secretory trap method, that circumvents these problems by making it possible to pre-screen for candidate axon guidance receptors

before generating lines of mutant mice, and to identify wiring defects readily through transgenic labelling of neurons expressing the trapped gene. The method has yielded numerous lines displaying many expression patterns, ranging from labelling throughout the nervous system to labelling of select axon tracts, creating a valuable resource for elucidating the normal pattern of axonal projections in the brain. In addition, screening for defects has revealed an axon guidance role for a vertebrate transmembrane semaphorin, *Sema6A*, and helped re-evaluate the function of the *EphA4* receptor tyrosine kinase. The PLAP secretory trap method thus provides a powerful way to access the biology and mechanisms of axonal guidance in the mammalian brain.

Rationale for a modified gene-trap strategy

To uncover mechanisms directly involved in axon guidance, we modified a gene-trapping procedure. In gene-trapping approaches, mutagenesis is performed in embryonic stem (ES) cells using a DNA construct that integrates at random in the mouse genome¹⁰. When the vector inserts into an intron, a fusion transcript is created through the splicing of the 'trapped' gene's upstream regions to vector sequences, in our case a 'secretory trap' form of β -geo (a fusion between β -galactosidase and neomycin phosphotransferase¹¹) that markedly enriches for insertions in candidate ligands and receptors^{12,13} (Fig. 1). The gene is easily identified by 5' rapid amplification of complementary DNA ends (RACE) and sequencing^{14,15}.

To help identify subtle wiring defects against a background of normal axonal projections, we added a second marker, human placental alkaline phosphatase (PLAP) and an internal ribosome entry site (IRES)^{16,17} (Fig. 1a). PLAP is a glycosyl phosphatidylinositol (GPI)-linked cell-surface protein that, when expressed transgenically in neurons, labels axons completely^{18,19}. When a gene is trapped, the endogenous promoter and enhancer elements direct production of a bicistronic transcript that encodes two proteins: the β -geo fusion to the endogenous protein and the PLAP protein, which is translated independently using the IRES (Fig. 1a).

The PLAP vector traps axon guidance molecules

a

ATG
1
S
2
SA
TM
 β -geo
TGA
ATG
IRES
PLAP
TGA
pA
3

Endogenous promoter

Secretory trap vector

b

X-gal

s
d
m
f
Cell bodies

c

PLAP

m
s
Axons

d

Trap secreted proteins in ES cells

Sequence and select lines for analysis

Assess phenotypes

Electroporate ES cells

Replica plate G418^r colonies

Screen for X-gal staining

5' RACE, sequence

BLAST search, prioritize lines for generating mice

Establish lines by blastocyst injection

Intercross heterozygous stock mice

Assess viability

Identify labelled axons

Look for wiring defects

Twenty-seven per cent of the trapped genes are new, and many of these also define molecularly distinct tracts (Fig. 3). *LST16*, which encodes a transmembrane protein with leucine-rich and immunoglobulin repeats (Fig. 3a), labels the habenula, piriform cortex and barrel cortex (Fig. 3b, c). *KST37* encodes a nidogen/plexin homology transmembrane protein (Fig. 3d) that is expressed in peripherally projecting axons (Fig. 1c), as well as in the fimbria and stripes of Purkinje cells in the cerebellum (Fig. 3e, f). *KST223* is weakly

* Known or likely axon guidance molecules.

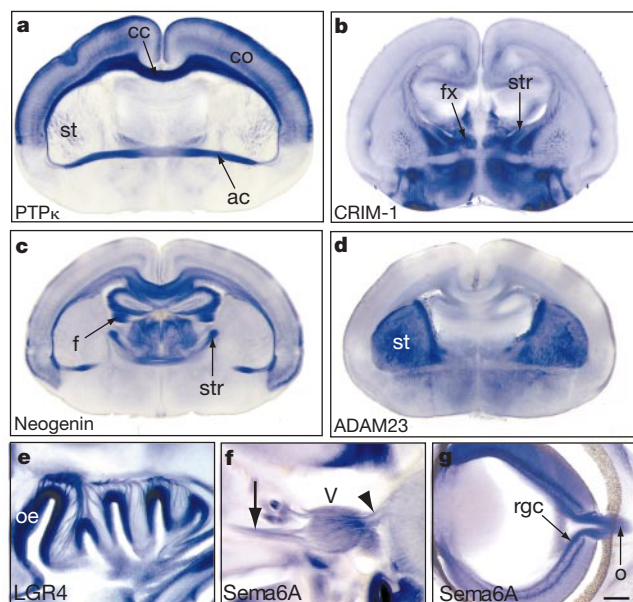


Figure 2 A survey of axonal populations labelled by PLAP. **a–d**, P0 brains, coronal sections, dorsal is up. ac, anterior commissure; cc, corpus callosum; co, cortex; f, fimbria; fx, fornix; st, striatum; str, stria terminalis. **a**, PTPK homozygote; **b**, CRIM-1 heterozygote; **c**, neogenin homozygote; **d**, ADAM23 homozygote. **e**, LGR4 heterozygote, E15.5, sagittal section. oe, olfactory epithelium. **f**, Sema6A homozygote, E12.5, sagittal section. Both peripheral (arrow) and central (arrowhead) projections of the trigeminal ganglion neurons (V) are labelled. **g**, Sema6A heterozygote, E15.5, coronal section. o, optic nerve; rgc, retinal ganglion cells. Scale bars, 500 μ m (**a–d**); 106 μ m (**e**); 225 μ m (**f**); 70 μ m (**g**).

related over much of its sequence to the interleukin-17 (IL-17) receptor (Fig. 3g), and shows remarkably specific expression in the fasciculus retroflexus (Fig. 3h) and in a restricted region of the inferior colliculus and cerebellum (Fig. 3i).

Axon guidance phenotypes in PLAP gene-trap mutants

Insertion of the PLAP vector induces either null or severe hypomorphic mutations¹³ that can be analysed at several levels. First, the mutants are assessed for lethality and gross defects, such as ataxia. Neurological defects have been observed in two mutants: *ADAM23* (ref. 20), in which homozygotes display tremor and ataxia¹³; and *EphA4*, in which homozygotes have an abnormal gait (see below). Second, lines that survive to at least E11.5 are screened systematically for wiring defects, even if the line is fully viable, because axon guidance molecules are not necessarily expected to cause overt phenotypes (for example see refs 8, 9). Eventually, viable lines can also be screened for defects in locomotion or behaviours such as learning and memory.

The anatomical screen has already revealed axon guidance defects in two mutant lines, *Sema6A* and *EphA4*. Preliminary analysis of these two mutants confirms, first, that we are able to identify new axon guidance phenotypes, and second, that characterization of PLAP staining can extend the analysis of genes that have already been knocked out.

Thalamocortical axon defects in *Sema6A* mutants

Semaphorins comprise a large family of transmembrane and secreted proteins that so far have been defined as guidance signals (for review see ref. 3). *Sema6A* belongs to a subfamily characterized by an extracellular semaphorin domain, a transmembrane domain

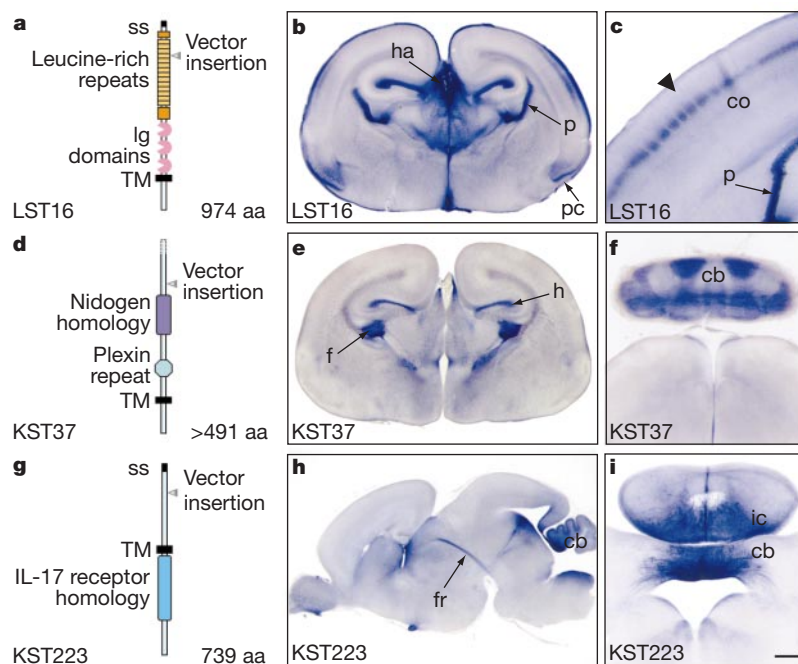


Figure 3 A sample of new predicted transmembrane proteins and their expression. All expression patterns are at P0 except for **c**, which is at P8. **a**, LST16 is a predicted protein of 974 amino acids (aa) comprising an N-terminal signal sequence (ss), 14 leucine-rich repeats (bracketed by N- and C-terminal leucine-rich repeat motifs), 3 immunoglobulin (Ig) domains, a single transmembrane domain and a cytoplasmic domain of 167 amino acids. It is 64% identical to another gene, KIAA0806 (GenBank accession no. XM_001307). **b**, LST16 homozygote, coronal section. ha, habenula; p, pia; pc, piriform cortex. **c**, LST16 heterozygote, coronal section. co, somatosensory cortex; arrowhead indicates barrels. **d**, KST37 is a predicted protein of more than 491 amino acids with a

nidogen-related domain, a plexin repeat, a transmembrane domain and a cytoplasmic domain of 54 amino acids (the extreme N terminus has not been determined), and is 55% identical over its length to tumour endothelial marker 7 (ref. 37). **e**, KST37 homozygote, coronal section. f, fimbria; h, hippocampus. **f**, KST37 homozygote, transverse section. cb, cerebellum. **g**, KST223 is a predicted protein of 739 amino acids that is weakly related to the IL-17 receptor, particularly in the cytoplasmic domain. **h**, KST223 heterozygote, sagittal section. fr, fasciculus retroflexus. **i**, KST223 homozygote, transverse section. ic, inferior colliculus. Scale bar, 500 μ m (**b**, **e**); 205 μ m (**c**); 275 μ m (**f**); 700 μ m (**h**), and 400 μ m (**i**).

and a long cytoplasmic tail^{21,22}. Members of this class, including *Sema6A*, can repel sympathetic and dorsal root ganglion axons *in vitro*^{22,23}, consistent with a traditional role as guidance signals. However, the length of the cytoplasmic tail, which includes an Evi-binding site in *Sema6A*²⁴ and a Src-binding site in *Sema6B*²⁵, suggests that these semaphorins may also function as receptors.

We isolated an insertion in the *Sema6A* gene at amino-acid 473 in the semaphorin domain that completely abolishes wild-type *Sema6A* transcripts¹³. Homozygous mutant mice are viable and fertile and display no obvious behavioural or morphological phenotypes. Staining for β -gal activity duplicates the published *in situ* hybridization pattern for *Sema6A*²¹ (Fig. 4a; and data not shown). PLAP staining labels specific axonal populations including the cranial nerves and optic tract (Fig. 2f, g), dorsal root ganglion axons, and several tracts in the brain including the fasciculus retroflexus, stria medullaris, the anterior commissure and thalamo-cortical axons (Fig. 4; and data not shown).

Although most of the tracts examined appeared unchanged in homozygotes, proper development of the thalamocortical projection requires *Sema6A* function. PLAP staining in heterozygotes shows thalamocortical axons projecting normally through the thalamus in a well-ordered fan-shaped array, turning sharply through the internal capsule and sweeping up to the cortex^{26,27} (Fig. 4c, d). X-gal staining labels cell bodies of these neurons in the thalamus (Fig. 4b). As there is only weak expression in the cortex at this stage, and no strong PLAP stain in cortical layer 6, it is likely that few if any descending corticothalamic axons are stained (Fig. 4c). Strikingly, thalamocortical axons in homozygotes fail to turn up through the internal capsule and instead project down towards the amygdala region (Fig. 4g–i). This defect is fully penetrant ($n = 12$), but is specific to caudal thalamocortical axons; rostral projections appear normal in every animal ($n = 12$) (data not shown). Aberrant thalamocortical axons were confirmed by injection of the lipophilic axonal tracer, DiI, into the dorsal thalamus (Fig. 4f, i).

The fact that *Sema6A* is expressed in thalamocortical neurons and required for their axons to project properly suggests that the

guidance defect is cell-autonomous and that *Sema6A* is acting in these axons as a guidance receptor. *Sema6A* is expressed broadly both in the thalamus and amygdala at the time when these axons are growing (E12.5–E15.5, Fig. 4a), however, which makes it impossible to distinguish fully between autonomous and non-autonomous activities without further experimentation. Alternatively, by analogy to the function of the *Drosophila Sema1a* gene (which has a closely related extracellular, but not intracellular domain)²⁸, *Sema6A* may function by promoting defasciculation of thalamo-cortical axons in a paracrine fashion, thereby enabling them to respond to other cues; misrouting would then be secondary to a failure to defasciculate.

Revisiting defects in *EphA4* mutants

Null mutations generated by the gene targeting of *EphA4* (refs 29, 30) possess defects in the corticospinal tract (CST; see Fig. 5a) and anterior commissure. As *EphA4* expression was not detected previously in the CST, a model was proposed in which an ephrin ligand on the axons senses EphA4 on spinal cord cells surrounding the CST²⁹. This possibility was supported by evidence that transmembrane ephrin ligands can also have a receptor function^{31,32}.

We have revisited the issue of whether EphA4 must have a ligand function by analysing an *EphA4* mutant generated by inserting the PLAP vector. Like the targeted *EphA4* mutants^{29,30}, homozygous *EphA4* gene-trap mice exhibit a hopping kangaroo gait and have guidance defects in the CST and anterior commissure (Fig. 5e; and data not shown). Unexpectedly, however, we found evidence for *EphA4* expression in CST neurons themselves. First, X-gal staining was observed in layer 5 of the motor cortex (as well as other layers) (Fig. 5b). Second, CST axons exhibited PLAP staining along their length, including in the internal capsule, at the pyramidal decussation, and in the spinal cord, where the pattern was consistent with expression in CST axons and their collateral branches (Fig. 5c, d; and ref. 33). In homozygotes, PLAP staining revealed an abnormal projection in the dorsal midline at the pyramidal decussation, suggesting crossing defects of CST axons (data not shown). In the spinal cord, the dorsal funiculus appeared smaller and shifted dorsally; there was also abnormal midline crossing of processes, which may consist partially or entirely of collateral branches of the CST axons (Fig. 5e).

The CST defects observed here are identical to those described previously²⁹. The fact that CST neurons produce β -gal and PLAP is, however, unexpected. The low level of expression of *EphA4* in the CST was easily detectable using the sensitive X-gal and PLAP staining but might be difficult to detect specifically with antibodies or by *in situ* hybridization, possibly explaining the discrepancy between our results and those of Dottori *et al.*²⁸ Indeed, *in situ* hybridization analysis has revealed intermediate level expression of *EphA4* in the cortex in a pattern identical to that observed by X-gal staining in the gene-trap line³³. Thus, EphA4 may indeed act as a receptor in the CST, a model further supported by the analysis of EphA4-kinase-deficient mice³³.

Discussion

We have developed and initiated a large-scale screen for new axon guidance molecules in the developing mouse nervous system. The screen combines two previous methods: the 'secretory trap' technique, used to mutate genes encoding proteins with signal sequences^{12,13}; and the use of IRES vectors to drive expression of an axonal marker in transgenic animals, previously used in targeted insertions into known genes^{16,34}. Our initial studies show that the PLAP secretory trap vector provides a useful tool for identifying the functions of both known and unknown genes in brain wiring, because it effectively mutates genes of interest and simultaneously introduces transgenic markers, β -geo and PLAP, that faithfully label the cell bodies and the axons of neurons expressing the trapped gene, respectively. This approach makes it possible to scan the

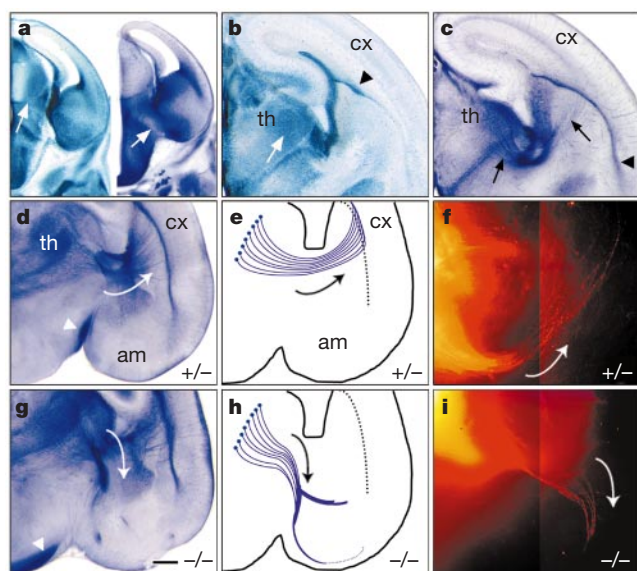


Figure 4 Thalamocortical axon misrouting in *Sema6A* mutants. Coronal sections, dorsal is up. **a**, *Sema6A* homozygote, E14.5, X-gal (hemisection on left; midline to left) and PLAP (right) staining. **b, c**, *Sema6A* heterozygote, P0, X-gal (**b**) and PLAP (**c**) staining. cx, cortex; th, thalamus; white arrow, dorsal thalamus. Arrowheads point to staining in a glia palisade. **d–i**, Thalamocortical projection (arrows) in *Sema6A* P0 heterozygotes (**d–f**) and homozygotes (**g–i**), shown with PLAP staining (**d, g**), in schematic form (**e, h**), or with DiI labelling (**f, i**; contrast increased selectively on right half). The optic tract (arrowheads) is unaffected. Scale bar, 350 μ m (**a–c**); 400 μ m (**d, g**); 800 μ m (**f, i**).

mammalian genome efficiently for axon guidance molecules.

The use of the PLAP marker in conjunction with the gene-trap vector solves, at least for cell-autonomous mechanisms, the most vexing problem facing a genetic screen for brain wiring mechanisms: how to examine only those axons most likely to be affected by the mutation among the many axons that project normally. As shown for *Sema6A* and *EphA4*, one can identify defects in the trajectories of axons expressing the trapped gene simply by comparing the wiring of PLAP-stained axons in heterozygous and homozygous embryos. The analysis of *Sema6A*, in particular, shows how PLAP staining can reveal previously unreported axon guidance phenotypes. The PLAP secretory trap method is likely to be most useful for identifying genes that encode axon guidance receptors or that have other cell-autonomous functions in guidance, as the axons that require the function of the gene are labelled with PLAP. This does not, however, preclude the discovery of non-autonomous functions, as defects in axons that do not express PLAP may sometimes be discernable.

Although we are using as our reference point the pattern of PLAP-stained axons in heterozygous rather than wild-type animals, this pattern will be identical to the wild-type pattern when the mutations are fully recessive (as is expected in most cases), and can be verified by comparison with known anatomy where available. In rare instances of haplo-insufficiency some defective projections may be present in the heterozygotes (for example, see ref. 30), but it is highly unlikely that they would be as severe as in homozygotes; therefore, the involvement of a trapped gene in axon guidance is unlikely to be obscured even in those cases.

Although the PLAP secretory trap screen in this study has already helped to identify a new axon guidance phenotype (in *Sema6A* mutants), and to characterize further a known phenotype (in *EphA4*

mutants), its full impact will come from its application on a large scale. Because the method does not rely on advance knowledge of the type of molecules that will be involved, axon guidance mechanisms in mammals can be identified without bias for the kinds of proteins or their expression levels. The PLAP secretory trap screen can also accelerate the enumeration of the various membrane proteins made by particular axonal tracts and therefore help build a molecular map of axonal projections. As the complement of surface proteins made by axons is defined, it might be possible to uncover a code of surface receptors that defines particular axonal trajectories, much as studies have defined a transcriptional code (the LIM code) for specifying motor neuron trajectories³⁵. Finally, the bank of ES cells and of mouse lines in which particular axonal populations are labelled with PLAP (which can be accessed through <http://www.genetrap.org>) provides a resource to help elucidate the normal wiring diagram of the mammalian brain. □

Methods

The PLAP vector

A fusion of the EMCV IRES¹⁶ to PLAP³⁶ was inserted downstream of β -geo in pGT1tm, a modification of pGT1.8TM^{12,13}. We introduced FRT sites flanking β -geo. Vectors in all three reading frames were generated and were named pGTOTMpf, pGT1TMpf and pGT2TMpf.

ES cell culture and sequencing

We cultured and screened E14 ES cells for X-gal staining as described¹². 5' RACE of ES cell clones was done as described¹³.

Mouse husbandry

F₁ heterozygous progeny (generated from crosses of chimaeras to C57Bl/6J (B6)) were backcrossed to B6. F₂ males and females were intercrossed to generate heterozygous and homozygous progeny for the screen. Genotypes were determined by quantitative dot blot hybridization using a *lacZ* probe¹³.

Histology

Tissues were fixed in 4% paraformaldehyde and embedded in 5% low-melt agarose. We collected 100- μ m sections using a vibrating microtome and air dried them onto Superfrost Plus slides (Fisher Scientific). β -gal activity was detected with 5-bromo-4-chloro-3-indolyl β -D-galactopyranoside (X-gal) under standard conditions, followed by Nuclear Fast Red counterstain (Vector Laboratories). We detected PLAP activity with AP staining buffer (0.1 mg ml⁻¹ 5-bromo-4-chloro-3-indolyl phosphate, 1 mg ml⁻¹ nitroblue tetrazolium in 100 mM Tris-HCl pH 9.5, 100 mM NaCl, 5 mM MgCl₂), after a 1-h incubation in PBS at 65 °C. PLAP stained tissue was dehydrated in methanol, cleared in benzyl alcohol:benzyl benzoate (1:2) and rehydrated.

Dil tracing

We fixed 300- μ m thick coronal sections in 4% paraformaldehyde and injected DiI (Molecular Probes) into the dorsal thalamus, before incubating them at 37 °C for 3 d to allow the DiI to diffuse down the axons.

Received 31 August 2000; accepted 5 January 2001.

1. Tessier-Lavigne, M. & Goodman, C. S. The molecular biology of axon guidance. *Science* **274**, 1123–1133 (1996).
2. Ebens, A. et al. Hepatocyte growth factor/scatter factor is an axonal chemoattractant and a neurotrophic factor for spinal motor neurons. *Neuron* **17**, 1157–1172 (1996).
3. Raper, J. A. Semaphorins and their receptors in vertebrates and invertebrates. *Curr. Opin. Neurobiol.* **10**, 88–94 (2000).
4. Hrabe de Angelis, M. et al. Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nature Genet.* **25**, 444–447 (2000).
5. Nolan, P. M. et al. A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nature Genet.* **25**, 440–443 (2000).
6. Kasarskis, A., Manova, K. & Anderson, K. V. A phenotype-based screen for embryonic lethal mutations in the mouse. *Proc. Natl Acad. Sci. USA* **95**, 7485–7490 (1998).
7. Martin, K. A. et al. Mutations disrupting neuronal connectivity in the *Drosophila* visual system. *Neuron* **14**, 229–240 (1995).
8. Giger, R. J. et al. Neuropilin-2 is required in vivo for selective axon guidance responses to secreted semaphorins. *Neuron* **25**, 29–41 (2000).
9. Chen, H. et al. Neuropilin-2 regulates the development of selective cranial and sensory nerves and hippocampal mossy fiber projections. *Neuron* **25**, 43–56 (2000).
10. Brennan, J. & Skarnes, W. C. Gene trapping in mouse embryonic stem cells. *Methods Mol. Biol.* **97**, 123–138 (1999).
11. Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.* **5**, 1513–1523 (1991).
12. Skarnes, W. C., Moss, J. E., Hurtley, S. M. & Beddington, R. S. Capturing genes encoding membrane and secreted proteins important for mouse development. *Proc. Natl Acad. Sci. USA* **92**, 6592–6596 (1995).

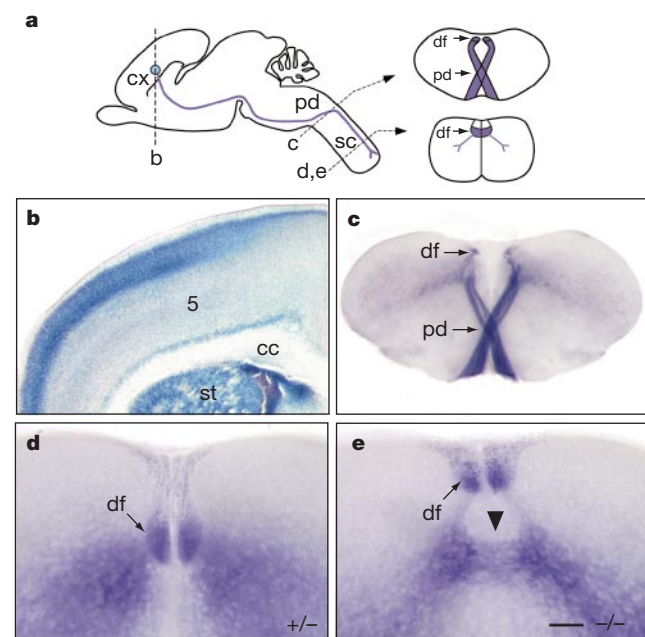


Figure 5 Axon guidance defects in *EphA4* mutants. **a**, The mouse corticospinal tract (CST): sagittal (left) and transverse (right) sections. Corticospinal neurons (blue circle) reside in the motor cortex (cx) and send their axons (purple line) across the midline and dorsally in the pyramidal decussation (pd, top right), projecting through the spinal cord in the dorsal funiculus (df, bottom right), with collateral branches extending into the grey matter. Dashed lines indicate coronal (**b**) and transverse (**c–e**) planes of section. **b**, *EphA4* heterozygote, P4. X-gal staining. 5, layer 5 of motor cortex; cc, corpus callosum; st, striatum. **c, d**, *EphA4* heterozygote, P4. PLAP staining in pyramidal decussation (**c**) and in dorsal funiculus (**d**). **e**, *EphA4* homozygote, P4. Arrowhead indicates abnormal crossing of midline in spinal cord. Scale bar, 315 μ m (**b**); 360 μ m (**c**), 60 μ m (**d, e**).

13. Mitchell, K. J. *et al.* Functional analysis of secreted and transmembrane proteins in mouse development. *Nature Genet.* (submitted).
14. Frohman, M. A., Dush, M. K. & Martin, G. R. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proc. Natl Acad. Sci. USA* **85**, 8998–9002 (1988).
15. Townley, D. J., Avery, B. J., Rosen, B. & Skarnes, W. C. Rapid sequence analysis of gene trap integrations to generate a resource of insertional mutations in mice. *Genome Res.* **7**, 293–298 (1997).
16. Mombaerts, P. *et al.* Visualizing an olfactory sensory map. *Cell* **87**, 675–686 (1996).
17. Mountford, P. S. & Smith, A. G. Internal ribosome entry sites and dicistronic RNAs in mammalian transgenesis. *Trends Genet.* **11**, 179–184 (1995).
18. Gustincich, S., Feigenspan, A., Wu, D. K., Koopman, L. J. & Raviola, E. Control of dopamine release in the retina: a transgenic approach to neural networks. *Neuron* **18**, 723–736 (1997).
19. Fields-Berry, S. C., Halliday, A. L. & Cepko, C. L. A recombinant retrovirus encoding alkaline phosphatase confirms clonal boundary assignment in lineage analysis of murine retina. *Proc. Natl Acad. Sci. USA* **89**, 693–697 (1992).
20. Sagane, K., Yamazaki, K., Mizui, Y. & Tanaka, I. Cloning and chromosomal mapping of mouse ADAM11, ADAM22 and ADAM23. *Gene* **236**, 79–86 (1999).
21. Zhou, L. *et al.* Cloning and expression of a novel murine semaphorin with structural similarity to insect semaphorin I. *Mol. Cell. Neurosci.* **9**, 26–41 (1997).
22. Kikuchi, K. *et al.* Cloning and characterization of a novel class VI semaphorin, semaphorin Y. *Mol. Cell. Neurosci.* **13**, 9–23 (1999).
23. Xu, X. M. *et al.* The transmembrane protein semaphorin 6A repels embryonic sympathetic axons. *J. Neurosci.* **20**, 2638–2648 (2000).
24. Klostermann, A., Lutz, B., Gertler, F. & Behl, C. The orthologous human and murine semaphorin 6A-1 proteins (SEMA6A-1/Sema6A-1) bind to the Enabled/Vasodilator-stimulated Phosphoprotein-like Protein (EVL) via a novel carboxyterminal Zyxin-like domain. *J. Biol. Chem.* **275**, 39467–39473 (2000).
25. Eckhardt, F. *et al.* A novel transmembrane semaphorin can bind c-src. *Mol. Cell. Neurosci.* **9**, 409–419 (1997).
26. Auladell, C., Pérez-Sust, P., Supèr, H. & Soriano, E. The early development of thalamocortical and corticothalamic projections in the mouse. *Anat. Embryol.* **201**, 169–179 (2000).
27. Molnár, Z., Adams, R. & Blakemore, C. Mechanisms underlying the early establishment of thalamocortical connections in the rat. *J. Neurosci.* **18**, 5723–5745 (1998).
28. Yu, H. H., Araj, H. H., Ralls, S. A. & Kolodkin, A. L. The transmembrane Semaphorin Sema I is required in *Drosophila* for embryonic motor and CNS axon guidance. *Neuron* **20**, 207–220 (1998).
29. Dottori, M. *et al.* EphA4 (Sek1) receptor tyrosine kinase is required for the development of the corticospinal tract. *Proc. Natl Acad. Sci. USA* **95**, 13248–13253 (1998).
30. Helmbacher, F., Schneider-Maunoury, S., Topilko, P., Tiet, L. & Charnay, P. Targeting of the EphA4 tyrosine kinase receptor affects dorsal/ventral pathfinding of limb motor axons. *Development* **127**, 3313–3324 (2000).
31. Henkemeyer, M. *et al.* Nuk controls pathfinding of commissural axons in the mammalian central nervous system. *Cell* **86**, 35–46 (1996).
32. Holland, S. J. *et al.* Bidirectional signalling through the EPH-family receptor Nuk and its transmembrane ligands. *Nature* **383**, 722–725 (1996).
33. Kullander, K. *et al.* Kinase-dependent and kinase independent functions of EphA4 receptors in major axon tract formation in vivo. *Neuron* **29**, 73–84 (2001).
34. Wang, F., Nemes, A., Mendelsohn, M. & Axel, R. Odorant receptors govern the formation of a precise topographic map. *Cell* **93**, 47–60 (1998).
35. Tsuchida, T. *et al.* Topographic organization of embryonic motor neurons defined by expression of LIM homeobox genes. *Cell* **79**, 957–970 (1994).
36. Gunning, P., Leavitt, J., Muscat, G., Ng, S. Y. & Kedes, L. A human beta-actin expression vector system directs high-level accumulation of antisense transcripts. *Proc. Natl Acad. Sci. USA* **84**, 4831–4835 (1987).
37. St Croix, B. *et al.* Genes expressed in human tumor endothelium. *Science* **289**, 1197–1202 (2000).

Acknowledgements

We thank R. Klein and colleagues for helpful discussions and sharing their results on *EphA4* mutants; S. McConnell, A. Chédotal, J. Rubenstein and members of the Rubenstein laboratory for helpful discussions on mouse neuroanatomy; A. Smith, P. Mombaerts and T. Vogt for reagents; and P. Tate, P. Wakenight and J. Mak for technical support. Funding for this project was provided by grants to M.T.L. and W.C.S. from the NIMH, and to W.C.S. from the NICHD. M.T.L. was also supported by a 1999 Sandler Award in Basic Sciences, the Howard Hughes Medical Institute, and the HHMI Research Resources Program grant to the UCSF School of Medicine. W.C.S. was a 1998 Searle Scholar. P.A.L. is a Howard Hughes Medical Institute Fellow of the Life Sciences Research Foundation, K.J.M. was supported by a fellowship from the Jane Coffin Childs Memorial Fund, and L.V.G. by a fellowship from the Helen Hay Whitney Foundation. X.L. is a postdoctoral associate and M.T.L. an Investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to M.T.-L. (e-mail: marctl@itsa.ucsf.edu).

Observation of orbital waves as elementary excitations in a solid

E. Saitoh*, S. Okamoto†, K. T. Takahashi*, K. Tobe*, K. Yamamoto*, T. Kimura*, S. Ishihara‡§, S. Maekawa† & Y. Tokura*‡

* Department of Applied Physics, University of Tokyo, Tokyo 113-8656, Japan

† Institute for Materials Research, Tohoku University, Sendai 980-8577, Japan

‡ Joint Research Center for Atom Technology (JRCAT)

and Correlated Electron Research Center (CERC), Tsukuba 305-0046, Japan

A basic concept in solid-state physics is that when some kind of symmetry in a solid is spontaneously broken, collective excitations will arise¹. For example, phonons are the collective excitations corresponding to lattice vibrations in a crystal, and magnons correspond to spin waves in a magnetically ordered compound. Modulations in the relative shape of the electronic clouds in an orbitally ordered state^{2–9} could in principle give rise to orbital waves, or ‘orbitons’, but this type of elementary excitation has yet to be observed experimentally. Systems in which the electrons are strongly correlated—such as high-temperature superconductors and manganites exhibiting colossal magnetoresistivity—are promising candidates for supporting orbital waves, because they contain transition-metal ions in which the orbital degree of freedom is important^{10,11}. Orbitally ordered states have been found in several transition-metal compounds^{12,13}, and orbitons have been predicted theoretically for LaMnO₃ (refs 4, 5). Here we report experimental evidence for orbitons in LaMnO₃, using Raman scattering measurements. We perform a model calculation of orbiton resonances which provides a good fit to the experimental data.

In the perovskite LaMnO₃, all of the Mn ions are trivalent with four electrons in 3d orbitals. Because of the crystal-field potential from surrounding oxygen ions and the strong Hund coupling, the electronic configuration of a Mn³⁺ is $t_{2g}^3 e_g^1$ (spin quantum number $S = 2$). Therefore, one electron occupies one of the doubly degenerate e_g orbitals, and the ion has both spin and orbital degrees of freedom^{14–16}. This latter degree of freedom is spontaneously frozen by the real-space ordering of the e_g orbitals accompanied by a Jahn–Teller (JT) type lattice distortion of the MnO₆ octahedra¹⁷; as shown in the inset of Fig. 1a, the $d_{3x^2-y^2}$ and $d_{3y^2-z^2}$ orbitals for e_g electrons are alternately ordered in the x – y plane (so called C-type orbital ordering) below the orbital ordering temperature $T_{OO} \approx 780$ K (ref. 13). In addition, the spin ordering appears below $T_N \approx 140$ K (refs 18, 19) where spins are aligned ferromagnetically in the x – y plane, and antiferromagnetically in the z direction (so-called A-type antiferromagnetic (AF) ordering). We address here the elementary excitations in the orbitally ordered state of LaMnO₃.

To describe the orbital degree of freedom theoretically, we introduce the pseudo-spin $\mathbf{T}$ (quantum number $T = \frac{1}{2}$) which represents two possible choices of the occupied e_g orbital at a Mn ion. The quantum number $T_z = +(-)\frac{1}{2}$ corresponds to the state where the $d_{3z^2-y^2}$ ($d_{3x^2-y^2}$) orbital is occupied by an electron. In a solid, there are interactions between pseudo-spins at different ions which give rise to an orbital order–disorder phase transition²⁰. In the orbitally ordered state, a macroscopic classical variable caused by the pseudo-spin—that is, the orbital order parameter—becomes finite. The order parameter in orbitally ordered LaMnO₃ is identified as $\mathbf{M}(\mathbf{k}) = (1/N)\sum_i e^{i\mathbf{k}\cdot\mathbf{r}_i} \langle \mathbf{T}_i \rangle = ((\sqrt{3}/4)\delta_{\mathbf{k}=(\pi,\pi,0)}, 0, -(1/4)\delta_{\mathbf{k}=(0,0,0)})$, where $\langle \dots \rangle$ indicates the thermal average, $\mathbf{r}_i$ is a position of the i th Mn ion and N is the number of ions. Consider the excitations of the orbital sector in the orbitally ordered state. Such an excitation corresponds to a change of the occupied orbital, that is, to a change

in the shape of the electron cloud. This is represented by a deviation of $\mathbf{T}_i$ from $\langle \mathbf{T}_i \rangle$.

Once the orbital excitation occurs at a site, it propagates in a solid through the interactions between orbitals at different ions, as does the spin excitation in a magnetically ordered state. This collective excitation and its quantized object are termed ‘orbital wave’ and ‘orbiton’ respectively. A schematic illustration is presented in Fig. 1b. The energy, dispersion and symmetry of the orbital wave are calculated from a model describing the interactions between orbitals. Here, considering the fact that LaMnO₃ is a Mott insulator and that the Hund coupling in a Mn ion is strong, the important interaction between orbitals is caused by the exchange of electrons between Mn ions. The model hamiltonian is^{4,5,16}:

$$H = -2 \sum_{\langle ij \rangle} \left\{ J_1 K_1 \left(\frac{1}{4} - \tau_i^l \tau_j^l \right) + J_2 K_2 \left(\frac{3}{4} + \tau_i^l \tau_j^l + \tau_i^l + \tau_j^l \right) \right\} \quad (1)$$

where $\tau_i^l = \cos(2\pi m_l/3)T_{iz} - \sin(2\pi m_l/3)T_{ix}$ with $(m_x, m_y, m_z) = (1, 2, 3)$, and where l denotes the direction of a bond connecting a site i and its nearest neighbouring site j . The quantities $K_1 = 1/4 - \mathbf{S}_i \cdot \mathbf{S}_j$ and $K_2 = 3/4 + \mathbf{S}_i \cdot \mathbf{S}_j$ involve $\mathbf{S}_i$ the spin operator

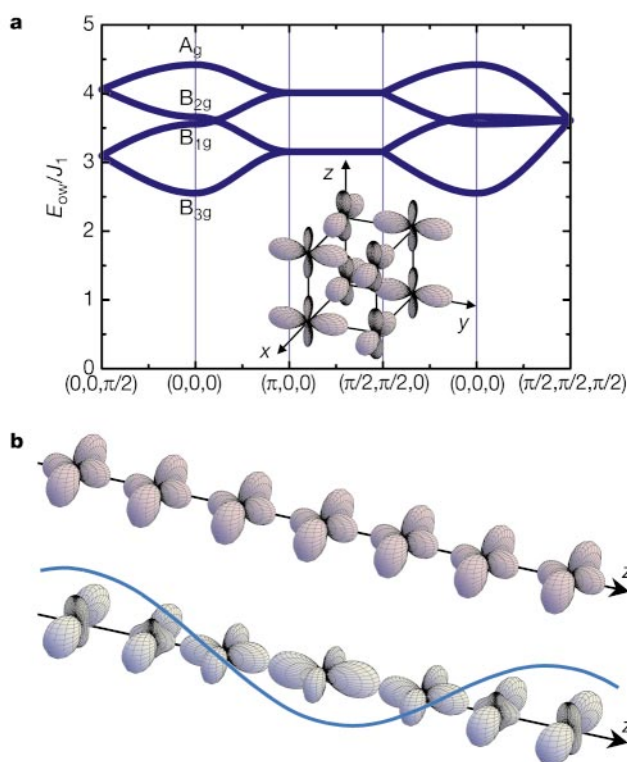


Figure 1 Dispersion relation and schematic illustration of an orbital wave. **a**, Theoretical results for the dispersion relation of the orbital wave in the C-type orbitally ordered state with the A-type AF structure (see text). The parameter values in equation (1) are chosen to be $J_2/J_1 = 0.35$. E_{OW} denotes energy of the orbital wave. In addition to the hamiltonian, equation (1), the interaction between the pseudo-spin $\mathbf{T}_i$ and the static Jahn–Teller-type lattice distortion in a MnO₆ octahedron is introduced. This is described by the hamiltonian $H_{JT} = g\sum_i (Q_{iz}T_{iz} + Q_{ix}T_{ix})$, where Q_{iz} and Q_{ix} are the two normal coordinates of the displacements of the O ions surrounding a Mn ion at site i , and g is a coupling constant. We choose $(1/N)\sum_i e^{i\mathbf{k}\cdot\mathbf{r}_i} \langle Q_{iz}, Q_{ix} \rangle = Q((\sqrt{3}/2)\delta_{\mathbf{k}=(\pi,\pi,0)}, -\frac{1}{2}\delta_{\mathbf{k}=(0,0,0)})$ and $gQ/J_1 = 0.7$. This interaction adds an energy $(0.8gQ)$ to the centre of gravity of the dispersion relation. The Brillouin zone for the tetragonal lattice with the D_{4h} symmetry is adopted. A schematic picture of the orbitally ordered state (electron cloud) in LaMnO₃ is shown in the inset. **b**, Schematic illustration of the orbital wave. Top, the electron cloud in the ground state. Bottom, a snapshot of the electron cloud when an orbital wave is excited. The wavevector of the orbital wave is chosen to be $(0, 0, 6\pi/a)$, with a being the bond length between the nearest neighbouring Mn ions.

§ Present address: Department of Applied Physics, University of Tokyo, Tokyo 113-8656, Japan.

with spin quantum number $S = 1/2$. Two terms in H correspond to the different exchange processes and (J_1, J_2) are amplitudes of these interactions. It has been shown^{4,5} that this model can well describe the actual situation of the spin and orbital states in LaMnO_3 . The exchange of virtual phonons also causes the interaction between orbitals. As it is estimated to be smaller than J_1 and has a similar form to equation (1), this contribution is included in the parameter $J_1 K_1$ of the equation.

The dispersion relation is calculated by applying the Holstein–Primakoff theory to equation (1) for the hypothetical tetragonal (D_{4h}) lattice of LaMnO_3 , and is presented in Fig. 1a. We assume that spin and lattice are frozen, as their dynamics are expected to be slower than that of the orbitals in LaMnO_3 . Then, the effect of the static JT-type lattice distortion in a MnO_6 octahedron is introduced, which adds some energy to the centre of gravity of the orbital dispersion described by equation (1). At the Γ point, the four modes of the orbital wave have A_g, B_{2g}, B_{1g} and B_{3g} symmetries of the D_{4h} group. The band width of the orbital wave is of the order of the exchange interactions $J_{1(2)}$. A gap of the order of the band width is opened due to the anisotropy in the interactions between pseudo-spins.

Raman scattering has been used to identify the orbital waves at the Γ point. We have measured the Raman scattering spectra for a detwinned single crystal of LaMnO_3 . The detwinned crystal enables us to investigate the detailed polarization dependence. Raman scattering measurements were performed using a microscope spectroscopy system in the nearly backscattering configuration. The 514.5-nm line (2.4 eV) of an Ar^+ laser was used for excitation, and the laser power at the focus spot ($\sim 5 \mu\text{m}$ in diameter) was kept below 1 mW. The excitation energy corresponds to the energy scale of the charge-transfer interaction between the O 2p and the Mn 3d orbitals^{21,22}. The polarization configuration of the incident and the scattered light is represented by (x, y) and so on. As illustrated in the inset of Fig. 1a, x, y and z are taken to be the directions of Mn–Mn bonds and $x + y = x' \ (x - y = y')$. Spectra were measured in six polarization configurations: (x, x) , (x, y) , (x', x') , (x', y') , (z, z) and (z, x) .

In Figs 2a and b, we display the Raman scattering spectra at 9 K (the orbitally ordered ground state) in four polarization configurations for the LaMnO_3 single crystal. We first discuss briefly phonon modes activated by the JT lattice distortions. Peak structures observed in a lower-energy region ($\leq 100 \text{ meV}$) can be assigned to phonon scattering, as discussed previously²³. Among them, two intense modes can be observed in the (x, x) polarization spectrum at 60 and 75 meV, and another mode for the (z, z) polarization at 60 meV. These modes have been identified as the activated oxygen modes due to a cooperative JT distortion²⁴. Figure 2c shows the temperature variation of these phonon modes in the (x, x) polarization spectrum. With an increase in temperature, the spectral intensity decreases and then almost disappears near the structural phase transition at $T_{\text{JT}} \approx 750 \text{ K}$ (ref. 17). The observed temperature dependence of these phonon peaks confirms that these modes reflect the order parameter of the JT lattice distortion.

The most notable feature in the present Raman study is found in a higher-energy region ($\geq 100 \text{ meV}$). In the (x, x) configuration spectrum at 9 K (Fig. 2a), three characteristic broad peak structures are observed at 125, 145 and 160 meV. One broad peak is also observed at 160 meV in the (z, z) configuration (Fig. 2b). The peak features in the energy region above $\sim 100 \text{ meV}$ show strong polarization dependence, which is displayed in Figs 3a–c. The high-energy Raman bands become weak with increasing temperature, and almost disappear around T_{OO} (Fig. 2d). This temperature profile parallels that of the phonon modes in the lower-energy region.

Raman bands for transition-metal oxides at such a high-energy scale are usually assigned to multi-phonon scattering. However, these Raman bands in the present compound cannot be attributed

to simple multi-phonon scattering for the following reasons. (1) The Raman-shift energy of the high-energy Raman bands cannot be reproduced by any combination of the observed Raman and infrared phonon modes. For comparison, we plot in Figs 2a and b the (x, x) and (z, z) spectra, respectively, as a function of doubled Raman shift (blue lines). Although the multi-phonon Raman scattering energy may shift from the sum of the one-

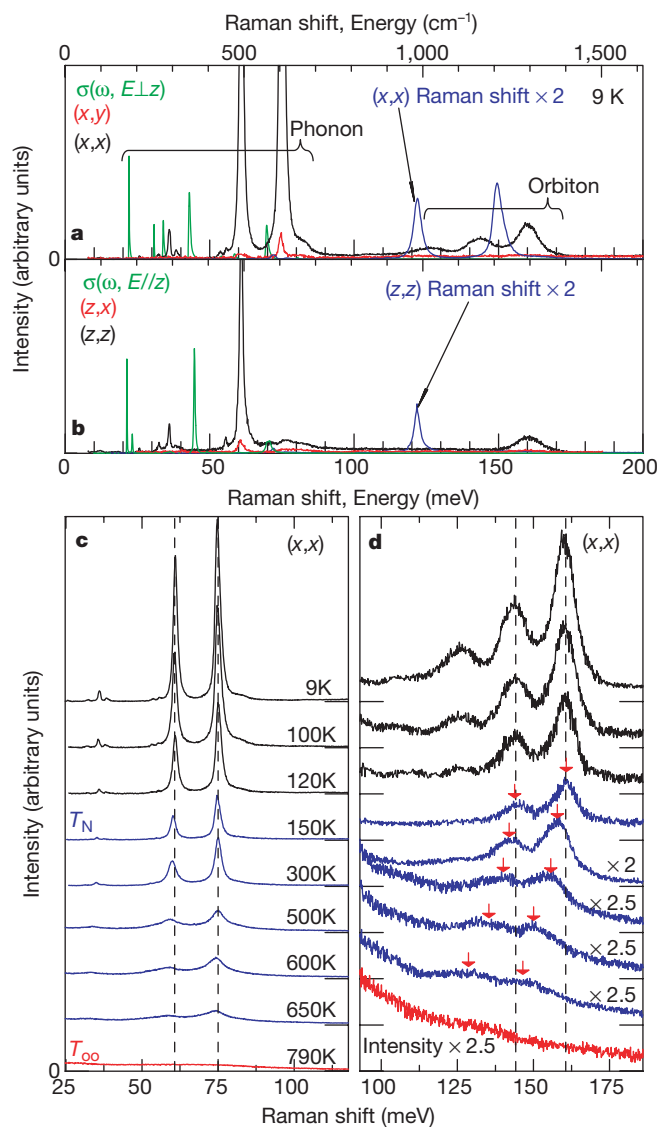


Figure 2 Experimental results for Raman scattering spectra of a detwinned LaMnO_3 crystal. **a, b**, Raman scattering spectra with four polarization configurations (black and red lines) and optical conductivity spectra (σ) (green lines) at 9 K. The crystal was obtained by the floating-zone method and a post-annealing procedure. The post-annealing was performed under an oxygen partial pressure of $3 \times 10^{-4} \text{ atm}$ and at a cooling rate of $\sim 10 \text{ K h}^{-1}$. To see the possible components of two-phonon excitations, we plot the (x, x) and (z, z) spectra as functions of the double Raman shift (blue lines). **c, d**, Temperature variation of Raman scattering spectra in the (x, x) configuration for the LaMnO_3 crystal. In **d**, the vertical axis (intensity) of the spectra above 300 K is stretched by a factor of 2 (300 K) or 2.5 (500–790 K). Black dashed lines are located at the peak positions of the respective Raman bands at 9 K. Although the peak positions of the phonon modes are almost independent of temperature (**c**), those of the Raman bands at higher energy (red arrows) shift appreciably with temperature above $T_{\text{N}} \approx 140 \text{ K}$ (**d**). This anomaly in the high-energy Raman band around T_{N} signals the close connection of the orbital state to the magnetic ordering, as expected from the original effective hamiltonian equation (1), and/or reflects the related change of the orbital order parameter¹³.

phonon energy to a limited extent due to the nonlinearity of lattice dynamics, such an effect cannot increase the multi-phonon energy. (We note that a weak phonon band has been reported around 650 cm^{-1} in previous studies on Raman spectra for $\text{LaMnO}_{3+\delta}$ (refs 25, 26). Such a phonon band is much weaker in the present Raman spectra, see Fig. 2a, but tends to grow with increasing oxygen contents or slightly doped Sr contents ($x \sim 0.06$) in $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$ where the high-energy Raman bands are rather suppressed.) (2) A substantial temperature-dependent energy shift of the peaks (up to 10 meV) is observed, in contrast to a negligible dependence of the observed phonon mode energies (see and compare Fig. 2d and c). (3) We also measured the spectra for RMnO_3 (where R is Nd or Gd). The decrease of the ionic radius of the R-site causes the suppression of the intensity of the phonon modes, but does not suppress that of the high-energy Raman bands. For these reasons, the observed Raman bands (for example, that at 160 meV) cannot be attributed to simple multi-phonon scattering. Charge excitations cannot be responsible for these Raman bands, because of the large charge gap energy ($\sim 1\text{ eV}$) in the ground state of LaMnO_3 (ref. 22). As for the two-magnon scattering, the A-type AF ordering (in-plane ferromagnetic) prohibits all the magnon scattering except for the (z, z) component.

The above unconventional high-energy excitations in the Raman scattering experiments are best interpreted in terms of the theoretical calculations of the orbital wave. The second-order interaction between electrons and photons creates an orbiton through the exchange of electrons. Consider a pair of the nearest neighbouring O and Mn ions in orbitally ordered LaMnO_3 . The incident photon excites an electron from the O $2p$ orbital to an unoccupied Mn e_g orbital. Via a second photon interaction, one of the two e_g electrons returns to the O ion. In the final scattering state, the orbital occupation of the Mn ion is changed, that is, a single orbiton is created by this two-photon process. The photon-induced exchange

of electrons between the nearest neighbouring Mn ions also excites orbitons²⁷. Both cross-sections for Raman scattering are estimated to be of the same order of magnitude as the two-magnon process in antiferromagnets. Electronic excitations between the crystal-field levels have been examined by Raman spectroscopy for $4f$ electron systems²⁸, where the local orbital excitations occur through the electronic transitions on a rare-earth ion. However, this is unlike the present case, showing the k -dispersion of the orbital excitation on the almost degenerate crystal-field levels.

We have calculated the Raman scattering spectra for several configurations of light polarization (Fig. 3d–f), and compared them with the experimental results (Fig. 3a–c). The actual small lattice distortions with monoclinic symmetry are taken into account²⁹. The peaks labelled A and B correspond to creation of single orbitons with A_g and B_{1g} symmetry (Fig. 1a) respectively. Peak C corresponds to the excitation at $\mathbf{k} = (\pi 00)$ in Fig. 1a, and becomes Raman active due to the monoclinic lattice distortion. (We estimate J_1 to be about 50 meV, from the analyses of the experimental spin wave spectrum in LaMnO_3 (ref. 30).) The theoretical results are in qualitatively good agreement with experiment. The following characteristics of the spectra provide strong evidence that the observed peaks originate from orbitons. (1) The strongest intensity of peak A for the (x, x) and (x', x') configurations originates from an interference effect in the in-phase A_g mode. (2) The weak peak C for the (x, x) and (x', x') configurations appears due to the small lattice distortion, as mentioned above. (3) The spectra for the (x, y) and (z, x) configurations are much weaker because, in the excitation process introduced above, an orbiton is not created for these configurations.

We expect that the present systematic theoretical and experimental studies of orbital waves and their quantized equivalents, orbitons, will also help us to understand better the thermal, electrical and magnetic properties of orbitally ordered systems. These properties are all strongly affected by the nature of the orbital state. It is certainly worthwhile to search for such orbital waves in other transition-metal compounds with orbital ordering, in order to ensure the general nature of this concept. $\square$

Received 9 November 2000; accepted 8 January 2001.

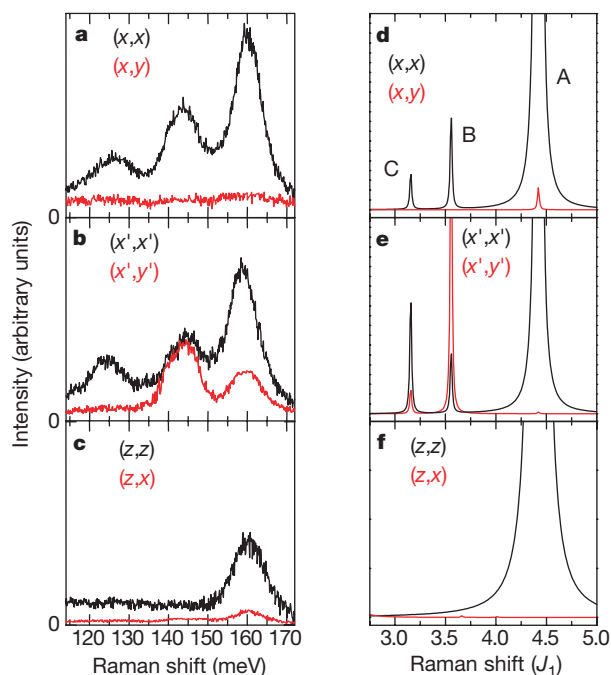


Figure 3 Comparison between the experimental and theoretical Raman scattering spectra for LaMnO_3 . **a–c**, Experimental Raman scattering spectra in several polarization configurations for a detwinned single crystal of LaMnO_3 at 9 K between 110 and 175 meV. **d–f**, Theoretical Raman scattering spectra due to orbitons in the C-type orbitally ordered state with the A-type AF structure. The displacements of the O ions observed in the monoclinic phase of LaMnO_3 are introduced in the calculation. J_1 is estimated to be about 50 meV from the analyses of the experimental spin wave spectrum in LaMnO_3 .

- Anderson, P. W. *Basic Notions of Condensed Matter Physics* (Benjamin-Cummings, London, 1984).
- Cyrot, M. & Lyon-Caen, C. Orbital superlattice in the degenerate Hubbard model. *J. Phys. (Paris)* **36**, 253–266 (1975).
- Komarov, A. G., Korovin, L. I. & Kudinov, E. K. Elementary excitation in systems with ‘frozen’ spin. *Sov. Phys. Solid State* **17**, 1531–1533 (1976).
- Ishihara, S., Inoue, J. & Maekawa, S. Electronic structure and effective Hamiltonian in perovskite Mn oxides. *Physica C* **263**, 130–133 (1996).
- Ishihara, S., Inoue, J. & Maekawa, S. Effective Hamiltonian in manganites: Study of the orbital and spin structures. *Phys. Rev. B* **55**, 8280–8286 (1997).
- Khalullin, G. & Oudovenko, V. Spin and orbital excitation spectrum in the Kugel-Khomskii model. *Phys. Rev. B* **56**, R14243–R14246 (1997).
- Feiner, L. E., Oleś, A. M. & Zaanen, J. Quantum disorder versus order-out-of-disorder in the Kugel-Khomskii model. *J. Phys. Condens. Matter* **10**, L555–L561 (1998).
- van den Brink, J., Stekelenburg, W., Khomskii, D. I., Sawatzky, G. A. & Kugel, K. I. Elementary excitations in the coupled spin-orbit model. *Phys. Rev. B* **58**, 10276–10282 (1998).
- Perebeinos, V. & Allen, P. B. Toward a theory of orbiton dispersion in LaMnO_3 . *Phys. Status Solidi B* **215**, 607–615 (1999).
- Tokura, Y. & Nagaosa, N. Orbital physics in transition-metal oxides. *Science* **288**, 462–468 (2000).
- Imada, M., Fujimori, A. & Tokura, Y. Metal-insulator transitions. *Rev. Mod. Phys.* **70**, 1039–1263 (1998).
- Ito, Y. & Akimitsu, J. Observation of orbital ordering in K_2CuF_4 . *J. Phys. Soc. Jpn* **40**, 1333–1338 (1976).
- Murakami, Y. *et al.* Resonant x-ray scattering from orbital ordering in LaMnO_3 . *Phys. Rev. Lett.* **81**, 582–585 (1998).
- Goodenough, J. B. Theory of the role of covalence in the perovskite-type manganites $[\text{La, M(II)}]\text{MnO}_3$. *Phys. Rev.* **100**, 564–573 (1955).
- Kanamori, J. Superexchange interaction and symmetry properties of electron orbitals. *J. Phys. Chem. Solids* **10**, 87–98 (1959).
- Kugel, K. I. & Khomskii, D. I. Crystal structure and magnetic properties of substances with orbital degeneracy. *Sov. Phys. JETP* **37**, 725–730 (1973).
- Rodríguez-Carvajal, J., Hennion, M., Moussa, F. & Moudén, A. H. Neutron diffraction study of the Jahn-Teller transition in stoichiometric LaMnO_3 . *Phys. Rev. B* **57**, R3189–R3192 (1998).
- Wollan, E. O. & Koehler, W. C. Neutron diffraction study of the magnetic properties of the series of perovskite-type compounds $[(1-x)\text{La, xCa}]\text{MnO}_3$. *Phys. Rev.* **100**, 545–563 (1955).
- Matsumoto, G. Study of $(\text{La}_{1-x}\text{Ca}_x)\text{MnO}_3$. I. Magnetic structure of LaMnO_3 . *J. Phys. Soc. Jpn* **29**, 606–615 (1970).

20. Kaplan, M. D. & Vekhter, B. G. *Cooperative Phenomena in Jahn-Teller Crystals* (Plenum, New York, 1995).
21. Saitoh, T. *et al.* Electronic structure of $\text{La}_{1-x}\text{Sr}_x\text{MnO}_3$ studied by photoemission and x-ray absorption spectroscopy. *Phys. Rev. B* **51**, 13942–13951 (1995).
22. Arima, T. & Tokura, Y. Optical study of electronic structure in perovskite-type RMO_3 ($\text{R} = \text{La}; \text{M} = \text{Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu}$). *J. Phys. Soc. Jpn* **64**, 2488–2501 (1995).
23. Iliev, M. N. *et al.* Raman spectroscopy of orthorhombic perovskite YMnO_3 and LaMnO_3 . *Phys. Rev. B* **57**, 2872–2877 (1998).
24. Yamamoto, K., Kimura, T., Ishikawa, T., Katsufuji, T. & Tokura, Y. Raman spectroscopy of the charge-orbital ordering in layered manganites. *Phys. Rev. B* **61**, 14706–14715 (2000).
25. Podobedov, V. B., Weber, A., Romero, D. B., Rice, J. P. & Drew, H. D. Effect of structural and magnetic transitions in $\text{La}_{1-x}\text{M}_x\text{MnO}_3$ ($\text{M} = \text{Sr, Ca}$) single crystals in Raman scattering. *Phys. Rev. B* **58**, 43–46 (1998).
26. Abrashev, M. V. *et al.* Comparative study of optical phonons in the rhombohedrally distorted perovskites LaAlO_3 and LaMnO_3 . *Phys. Rev. B* **59**, 4146–4153 (1999).
27. Inoue, J. *et al.* Raman scattering by orbital waves in perovskite LaMnO_3 . *Physica B* **237–238**, 51–53 (1997).
28. Schaack, G. in *Light Scattering in Solids VII* (eds Cardona, M. & Güntherodt, G.) 24–173 (Springer, Berlin, 2000).
29. Mitchell, J. F. *et al.* Structural phase diagram of $\text{La}_{1-x}\text{Sr}_x\text{MnO}_{3+\delta}$: Relationship to magnetic and transport properties. *Phys. Rev. B* **54**, 6172–6183 (1996).
30. Hirota, K., Kaneko, N., Nishizawa, A. & Endoh, Y. Two-dimensional planar ferromagnetic coupling in LaMnO_3 . *J. Phys. Soc. Jpn* **65**, 3736–3739 (1996).

Acknowledgements

We thank N. Nagaosa, T. Katsufuji, P. Prelovšek and S. E. Barnes for discussions. This work was supported in part by NEDO Japan, CREST Japan, the Science and Technology Special Coordination Fund for Promoting Science and Technology, and Grant-in-Aid for Scientific Research Priority Area from the Ministry of Education, Science and Culture of Japan. S.O. acknowledges the financial support of JSPS. Part of the numerical calculation was performed in the supercomputing facilities in IMR, Tohoku University.

Correspondence and requests for materials should be addressed to Y.T.
(e-mail: tokura@ap.t.u-tokyo.ac.jp).

Coherent branched flow in a two-dimensional electron gas

M. A. Topinka^{*}, B. J. LeRoy[†], R. M. Westervelt^{*†}, S. E. J. Shaw[‡],
R. Fleischmann[‡], E. J. Heller^{‡§}, K. D. Maranowski^{||} & A. C. Gossard^{||}

^{*} Division of Engineering and Applied Sciences, [†] Department of Physics,

[‡] Department of Chemistry and Chemical Biology, Harvard University,
Cambridge, Massachusetts 02138, USA

[§] Max-Planck-Institut für Strömungsforschung, Bunsenstrasse 10,
D-37073 Göttingen, Germany

^{||} Materials Department, University of California, Santa Barbara,
California 93106, USA

Semiconductor nanostructures based on two-dimensional electron gases (2DEGs) could form the basis of future devices for sensing, information processing and quantum computation. Although electron transport in 2DEG nanostructures has been well studied, and many remarkable phenomena have already been discovered (for example, weak localization, quantum chaos, universal conductance fluctuations^{1,2}), fundamental aspects of the electron flow through these structures have so far not been clarified. However, it has recently become possible to image current directly through 2DEG devices using scanning probe microscope techniques^{3–13}. Here, we use such a technique to observe electron flow through a narrow constriction in a 2DEG—a quantum point contact. The images show that the electron flow from the point contact forms narrow, branching strands instead of smoothly spreading fans. Our theoretical study of this flow indicates that this branching of current flux is due to focusing of the electron paths by ripples in the background potential. The strands are decorated by interference fringes separated by half the Fermi wavelength, indicating the persistence

of quantum mechanical phase coherence in the electron flow. These findings may have important implications for a better understanding of electron transport in 2DEGs and for the design of future nanostructure devices.

Images of electron flow from the quantum point contact (QPC) are obtained by raster scanning a negatively charged scanning probe microscope (SPM) tip above the surface of the device and simultaneously measuring the position-dependent conductance of the device. The negatively charged tip capacitively couples to the 2DEG, creating a depletion region that backscatters electron waves. When the tip is positioned over areas with high electron flow from the QPC, the conductance is decreased, whereas when the tip is over areas of relatively low electron flow the conductance is unmodified. By raster scanning the tip over the sample, a two-dimensional image of electron flow can be obtained.

The QPC sample is mounted in an atomic force microscope and cooled to liquid He temperatures. The QPC is formed in the 2DEG inside a GaAs/AlGaAs heterostructure by negatively biasing two gates on the surface—a negative potential on these gates creates two depletion regions that define a variable-width channel between them, as shown in Fig. 1a. The heterostructure for the devices used in this experiment was grown by molecular beam epitaxy on an n-type GaAs substrate. The 2DEG resides 57 nm below the surface with mobility $\mu = 1.0 \times 10^6 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$ and density $n = 4.5 \times 10^{11} \text{ cm}^{-2}$. These values of mobility and density correspond to a mean free path $l = 11 \text{ } \mu\text{m}$, Fermi wavelength $\lambda_F = 37 \text{ nm}$, and

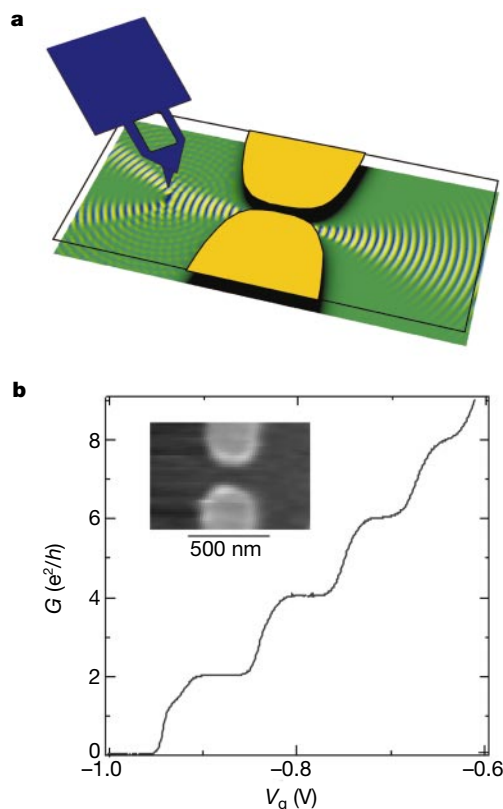


Figure 1 Experimental set-up. **a**, Schematic diagram of the experimental set-up used for imaging electron flow. The tip introduces a movable depletion region which scatters electron waves flowing from the quantum point contact (QPC). An image of electron flow is obtained by measuring the effect the tip has on QPC conductance as a function of tip position. Two ohmic contacts $\sim 1 \text{ } \mu\text{m}$ away from the QPC (not shown) allow the conductance of the QPC to be measured using an a.c. lock-in amplifier at 11 kHz. The root-mean-square voltage across the QPC, 0.2 mV, was chosen in order not to heat electrons significantly above the lattice temperature of 1.7 K. **b**, Conductance of the QPC used for Fig. 2b versus QPC width controlled by the gate voltage. Steps at integer multiples of $2e^2/h$ are clearly visible. The inset is a topographic AFM image of the QPC.

Fermi energy $E_F = 16$ meV. The conductance of the QPC, shown in Fig. 1b, increases as the width of the channel is increased (by changing the gate voltage V_g) and shows well defined conductance plateaus at integer multiples of the conductance quantum $2e^2/h$ (refs 1, 2).

Figure 2a and b shows images of electron flow from two different QPCs at the temperature 1.7 K; both QPCs are biased on the $G = 2e^2/h$ conductance plateau. Figure 2b shows the flow patterns on each side of a QPC and Fig. 2a shows a higher-resolution image of flow from one side of a different QPC. In both these images, the current leaves the point contact in a central lobe, as expected from an exact quantum-mechanical calculation of electron flow through an ideal QPC. Rather than continuing out as a smoothly widening fan, it quickly forks into several different paths and continues to branch off into ever smaller rivulets for the full width of the scan. This branching behaviour was observed in all of the 13 QPC exit patterns observed so far. Previously, there have been suggestions of an unexpected narrowness in observed flow from a QPC⁵, but until now, high-resolution, detailed images of electron flow from a QPC have been difficult to obtain and no observations or predictions for this type of strong branching behaviour have been published. The average electron flow reflected by the tip back through the QPC falls off approximately as $1/r^2$ with distance r from the the QPC.

In order to explore the experimentally observed current flow, we numerically calculated a set of conducting wavefunctions through a

QPC with disordered background potentials. Our model for the potential landscape incorporates known properties of 2DEGs in GaAs/AlGaAs heterostructures^{14,15} and replicates the physical parameters of the experimental system. To model the full in-plane potential, we consider two contributions. The first contribution is from the negatively charged gates that define the QPC, which we model using a smooth analytic potential that reproduces the quantized conductance steps characteristic of the QPC. The Fermi level remains constant inside the 2DEG, so that the propagation there is primarily affected by the second piece of the potential, the disordered background. There are two contributions to this background that we considered: the donor dopant atoms and unwanted impurities. For the impurities, distributed throughout the crystal structure, we take a random distribution in three dimensions and match the impurity atom density with a reasonable estimate for the physical sample ($1.25 \times 10^{15} \text{ cm}^{-3}$). The donors are located in a plane displaced from the 2DEG by 22 nm with a sheet density of $8 \times 10^{12} \text{ cm}^{-2}$; we estimate that half of the donors are ionized. Their distribution is random in space except for correlations limiting the maximum local density¹⁴. We use a $1/r^3$ potential at large distances for the impurities and donors, where r is the distance between a location in the plane of the 2DEG and the location of a given impurity or donor atom in the crystal. This $1/r^3$ dependence is the key feature of the full screened potential in a 2DEG from a point charge¹⁵.

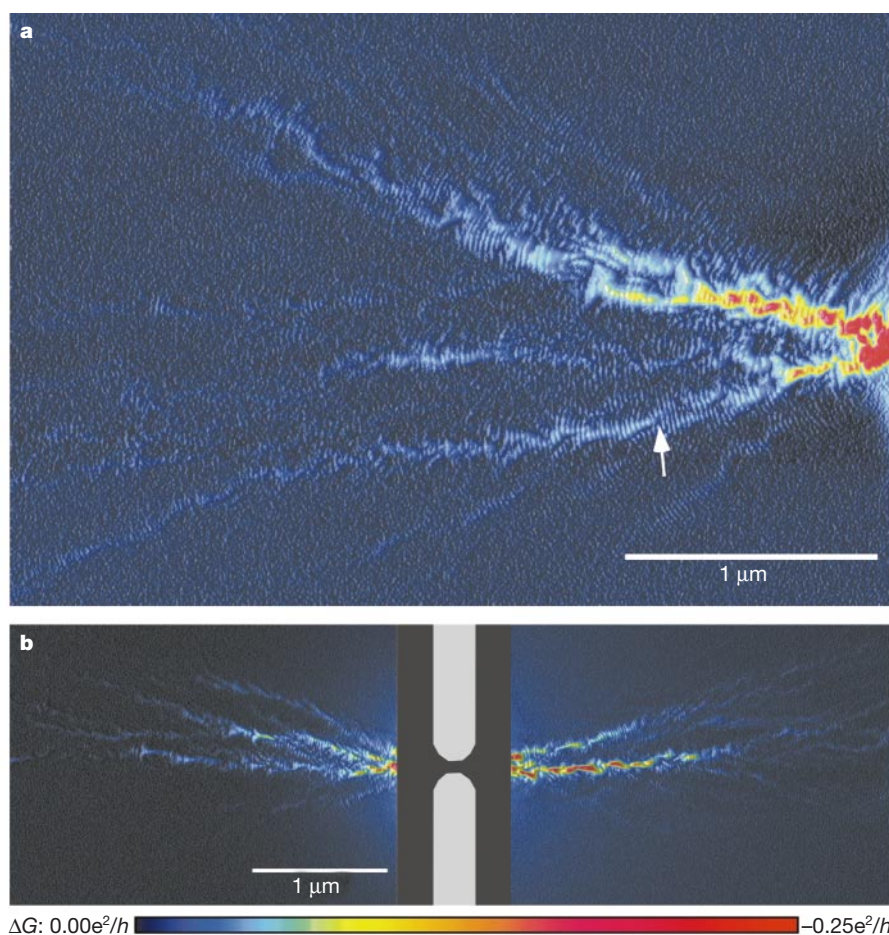


Figure 2 Experimental images of electron flow. **a**, Image of electron flow from one side of a QPC at $T = 1.7$ K, biased on the $G = 2e^2/h$ conductance step. Dark regions correspond to areas where the tip had little effect on QPC conductance, and hence are areas of low electron flow. The colour varies and the height in the scan increases with increasing electron flow. Narrow branching channels of electron flow are visible, and fringes spaced

by $\lambda_F/2$, half the Fermi wavelength, are seen to persist across the entire scan. **b**, Images of electron flow from both sides of a different QPC, again biased on the $G = 2e^2/h$ conductance step. The gated region in the centre was not scanned. Strong channelling and branching are again clearly visible. The white arrow points out one example of the formation of a cusp downstream from a dip in the potential.

Using the values for donor inhomogeneity and impurity density expected for the 2DEG sample in our model¹⁴, we calculated the mobility classically using the momentum relaxation time for an appropriate ensemble of trajectories and found it to match the measured value to within 10%. The distribution of energies in the disordered potential is very nearly gaussian with a standard deviation of about 8% E_F . The correlation length of the disordered potential is about 25 nm, that is, of the order of the wavelength.

In Fig. 3a we show a typical potential used in both classical and quantum simulations. Figure 3b and c shows the results of classical and quantum-mechanical flux density calculations in that potential which clearly show branched flow. The agreement between the classical and quantum results is very good, leading to the conclusion that the branched flow is essentially a classical phenomenon. Though there are occasional events where the flow is split by an impurity near the 2DEG, most of the bumps in the potential are well below the Fermi energy of the electrons. Locally, potential valleys act like lenses that focus the electron paths, albeit not perfectly, giving rise to a near-focal point known as a cusp. One example of such a cusp in the experimental images may be seen in Fig. 2a, as indicated by the arrow. The stronger branches surviving at large distances are the indirect result of passing over many hills and valleys (S.E.J.S. and E.J.H., manuscript in preparation), as shown in Fig. 3.

This classical phenomenon is robust, and it is seen in smooth potentials of disparate origin with features generally well below the energy of the scattered particles. The length scale for the formation of the channels is determined by the autocorrelation length of the potential. This regime of classical dynamics in a random potential

has not been studied in detail, falling as it does between the well explored extremes of gaussian white noise and widely spaced point scatterers. A detailed study of this phenomenon will appear elsewhere (S.E.J.S. and E.J.H., manuscript in preparation). Similar scattering, with the unexpected resulting strands of concentration of flux, might be expected in widely different contexts, such as sound propagation through the ocean, where it has recently been discussed¹⁶.

The experimental images in Fig. 2 show the position-dependent effect of the SPM tip on the conductance through the device. The charged tip induces an approximately lorentzian bump in the potential¹⁷ seen by electrons propagating through the system; we can simulate the experiments by adding a lorentzian to the potential at the tip position. Figure 4a shows the overall computed current flow through the device before the addition of a lorentzian. We compared the simulated flux in a small patch (Fig. 4b) with the conductance as a function of lorentzian position (Fig. 4c). This comparison confirms the relationship between flux and the images achieved by the experimental technique.

A striking feature of the experimental images (Fig. 2a and b) is the appearance of fringes oriented perpendicularly to electron flow and spaced by $\lambda_F/2$, half the Fermi wavelength. These fringes are caused by coherent constructive and destructive backscattering of electrons from the tip. In the region close to the point contact the fringes can be explained by multiple reflections of electron waves between the tip and the QPC gates. At farther distances two effects could suppress the fringes: phase decoherence and thermal broadening. In our sample the phase coherence length, $l_\phi = 18 \mu\text{m}$ ¹⁸, is much larger than our scan area and thus not a concern. The thermal length $l_{th} = \hbar 2\pi / m \lambda_F k_B T = 1.4 \mu\text{m}$ (where T is temperature, k_B is Boltzmann's constant, $\hbar$ is Planck's constant divided by 2π , m is the mass of the electron, and λ_F is the Fermi wavelength), the distance over which electrons differing in energy by $k_B T$ drift 1 radian out of phase, is relevant. The fringes in our experimental and theoretical images, however, persist up to this length and beyond.

The persistence of the fringes can be explained by an unusual source of coherent backscattering (S.E.J.S., R.F. and E.J.H., manu-

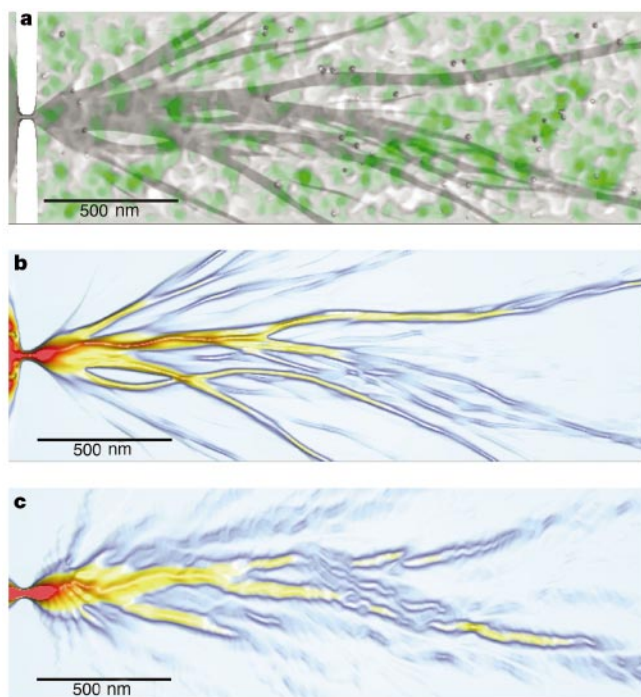


Figure 3 Calculated electron flow. **a**, Surface plot of the random potential for computed electron flow, including contributions from impurities, donors, and gates; green areas are low and white areas are high potential. The 'shadow' is cast by classical flux through the same potential. We note that the branched flux does not follow valleys in the potential. **b**, Classical and **c**, quantum-mechanical flux of electrons flowing through the potential in **a**. In the classical case, we followed the dynamics of an appropriate ensemble of classical trajectories and show the classical flux density. The quantum-mechanical results show the flux density of the transmitted wavefunction, coming through the point contact on the left. We note that both results show the same branching behaviour.

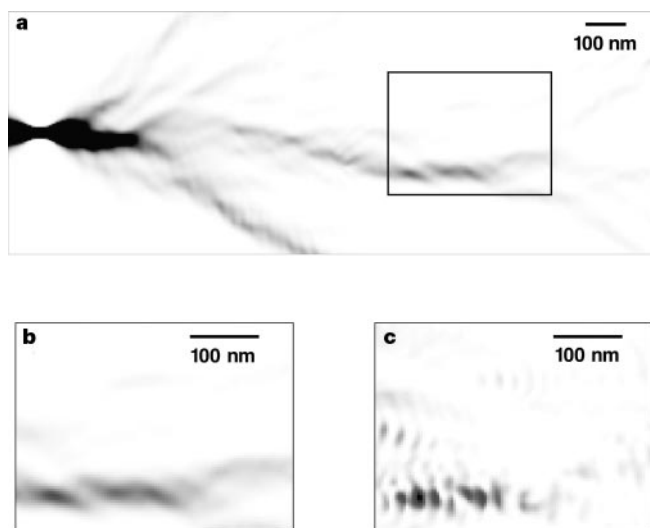


Figure 4 Calculated tip scan. **a**, Quantum-mechanical flux through a random potential. **b**, The flux from the boxed area in **a**. **c**, A raster scan of conductance as a function of SPM tip position in the same system as **a** and **b**. The conductance image in the model corresponds to the flux image, confirming our assertion that the experiment images electron flow. Additionally, the simulation **c** shows quantum fringes, as seen in the experiment. Though this simulation is at zero temperature, the fringes do survive thermal averaging.

script in preparation). The electron waves that are backscattered from the depletion region under the tip (located a distance r_{tip} from the QPC), interfere with a combination of the backscattered waves from all impurities in a ring covering the area $r_{\text{tip}} \pm l_{\text{th}}$ away from the QPC to produce alternating constructive and destructive total interference as r_{tip} is changed. As the tip is moved, the phase of the waves backscattered from the tip seen at the QPC varies as $2r_{\text{tip}}k$, where k is the wavevector. Importantly, this mechanism is resistant to thermal broadening, because as k is varied around k_F the phases of backscattered waves returning from impurities within the radius $r_{\text{tip}} \pm l_{\text{th}}$ vary in step with the phase from the tip (S.E.J.S., R.F. and E.J.H., manuscript in preparation) with the result that the fringes can easily survive. This mechanism requires phase-coherent transport over the round-trip distance to the most remote fringes, and may provide a new direct way to measure electron wave coherence length.

Another unusual feature of the fringes is that their separations are smaller than the width of the lorentzian perturbation used in making the measurements. The tip perturbation has an estimated half-width at half-maximum of ~ 60 nm, based on electrostatic simulations (ref. 17 and M.A.T. *et al.*, manuscript in preparation). Our numerical work, however, gives evidence that the relevant feature is the presence of a classically forbidden “depletion region” beneath the tip. The edge of this depletion region provides the backscattering needed to direct electrons back through the QPC. This result is in accord with experimental findings that images of electron flow are only observed when the voltage on the tip is sufficient to create a small depletion region in the 2DEG (M.A.T. *et al.*, manuscript in preparation). Moreover, backscattering reflection from a circular depletion region where paths reverse and lead back to the QPC comes only from a small zone on the depletion region.

The information about electron flow made available by this imaging technique has several important implications and possible future applications. The observation that electrons can flow through a 2DEG in narrow, branching channels may be important for experiments relying on the assumption of ‘ballistic’ flow of electrons for distances less than the mean free path. Such detailed images of electron flow may also prove valuable in future explorations of spin transport for spintronics and possibly allow novel quantum computer implementations. The surprising persistence of coherent fringes well past the thermal length may help provide additional insight into a broad range of coherent phenomena including universal conductance fluctuations, phase coherence, and weak localization. This imaging technique can provide much detail about coherent electron flow, and is an important tool in investigating the underlying physics as well as the future design of 2DEG nanostructures. □

Received 25 October 2000; accepted 2 January 2001.

1. Beenakker, C. W. J. & van Houten, H. in *Solid State Physics* (eds Ehrenreich, H. & Turnbull, D.) 44, 1–228 (Academic, New York, 1991).
2. Sohn, L. L., Kouwenhoven, L. P. & Schon, G. *Mesoscopic Electron Transport in Semiconductor Nanostructures* (Kluwer Academic, Boston, 1997).
3. Topinka, M. A. *et al.* Imaging coherent electron flow from a quantum point contact. *Science* **289**, 2323–2326 (2000).
4. Eriksson, M. A. *et al.* Cryogenic scanning probe characterization of semiconductor nanostructures. *Appl. Phys. Lett.* **69**, 671–673 (1996).
5. Crook, R., Smith, C. G., Simmons, M. Y. & Ritchie, D. A. Imaging cyclotron orbits and scattering sites in a high-mobility two-dimensional electron gas. *Phys. Rev. B* **62**, 5174–5178 (2000).
6. Crook, R. *et al.* Imaging diffraction-limited electronic collimation from a non-equilibrium one-dimensional ballistic constriction. *J. Phys. Cond. Matter* **12**, L167–L172 (2000).
7. McCormick, K. L. *et al.* Scanned potential microscopy of edge and bulk currents in the quantum Hall regime. *Phys. Rev. B* **59**, 4654–4657 (1999).
8. Zhitenev, N. B. *et al.* Imaging of localized electronic states in the quantum Hall regime. *Nature* **404**, 473–476 (2000).
9. Finkelstein, G., Glicofridis, P. I., Ashoori, R. C. & Shayegan, M. Topographic mapping of the quantum Hall liquid using a few-electron bubble. *Science* **289**, 90–94 (2000).
10. Manoharan, H. C., Lutz, C. P. & Eigler, D. M. Quantum mirages formed by coherent projection of electronic structure. *Nature* **403**, 512–515 (2000).
11. Jamneala, T., Madhavan, V., Chen, W. & Crommie, M. F. Scanning tunneling spectroscopy of transition-metal impurities at the surface of gold. *Phys. Rev. B* **61**, 9990–9993 (2000).

12. Yoo, M. J. *et al.* Scanning single-electron transistor microscopy: imaging individual charges. *Science* **276**, 579–582 (1997).
13. Gurevich, L., Canali, L. & Kouwenhoven, L. P. Scanning gate spectroscopy on nanoclusters. *Appl. Phys. Lett.* **76**, 384–386 (2000).
14. Grill, R. & Döhler, G. H. Effect of charged donor correlation and Wigner liquid formation on the transport properties of a two-dimensional electron gas in modulation delta-doped heterojunctions. *Phys. Rev. B* **59**, 10769–10777 (1999).
15. Davies, J. H. *The Physics of Low-Dimensional Semiconductors: An Introduction* (Cambridge Univ. Press, New York, 1998).
16. Wolfson, M. A. & Tomsovic, S. On the stability of long-range sound propagation through a structured ocean. At xxx.lanl.gov/abs/nlin.CD/0002030 (2000).
17. Eriksson, M. A. *et al.* Effect of a charged scanned probe microscope tip on a subsurface electron gas. *Superlattices Microstruct.* **20**, 435–440 (1996).
18. Althuler, B. L., Aronov, A. G. & Khmelnitsky, D. E. Effects of electron-electron collisions with small energy transfers on quantum localisation. *J. Phys. C* **15**, 7367–7386 (1982).

Acknowledgements

This work was supported in part at Harvard University by the Office of Naval Research/Augmentation Awards for Science and Engineering Research Training (ONR/AASERT), by ONR and by the National Science Foundation through grants for the Materials Research Science and Engineering Center and for the Institute for Theoretical Atomic and Molecular Physics at Harvard University and Smithsonian Astrophysical Observatory. Work at the University of California Santa Barbara was supported by the NSF Science and Technology Center QUEST.

Correspondence and requests for materials should be addressed to R.M.W. (e-mail: westervelt@deas.harvard.edu).

Strongly linked current flow in polycrystalline forms of the superconductor MgB_2

D. C. Larbalestier^{*†}, L. D. Cooley^{*}, M. O. Rikel^{*}, A. A. Polyanskii^{*}, J. Jiang^{*}, S. Patnaik^{*}, X. Y. Cai^{*}, D. M. Feldmann^{*}, A. Gurevich^{*}, A. A. Squitieri^{*}, M. T. Naus^{*}, C. B. Eom^{*†}, E. E. Hellstrom^{*†}, R. J. Cava[‡], K. A. Regan[‡], N. Rogado[‡], M. A. Hayward[‡], T. He[‡], J. S. Slusky[‡], P. Khalifah[‡], K. Inumaru[‡] & M. Haas[‡]

^{*} Applied Superconductivity Center, University of Wisconsin–Madison, 1500 Engineering Drive, Madison, Wisconsin 53706, USA

[†] Department of Materials Science and Engineering, University of Wisconsin–Madison, 1509 University Avenue, Madison, Wisconsin 53706, USA

[‡] Department of Chemistry and Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, USA

The discovery of superconductivity at 39 K in magnesium diboride¹, MgB_2 , raises many issues, a critical one being whether this material resembles a high-temperature copper oxide superconductor or a low-temperature metallic superconductor in terms of its behaviour in strong magnetic fields. Although the copper oxides exhibit very high transition temperatures, their in-field performance² is compromised by their large anisotropy, the result of which is to restrict high bulk current densities to a region much less than the full magnetic-field–temperature (H – T) space over which superconductivity is found. Moreover, the weak coupling across grain boundaries makes transport current densities in untextured polycrystalline samples low and strongly sensitive to magnetic field^{3,4}. Here we report that, despite the multiphase, untextured, microscale, subdivided nature of our MgB_2 samples, supercurrents flow throughout the material without exhibiting strong sensitivity to weak magnetic fields⁵. Our combined magnetization, magneto-optical, microscopy and X-ray investigations show that the supercurrent density is mostly determined by flux pinning, rather than by the grain boundary connectivity. Our results therefore suggest that this new superconductor class is not

compromized by weak-link problems, a conclusion of significance for practical applications if higher temperature analogues of this compound can be discovered.

The principal samples were synthesized by direct reaction of bright Mg flakes (Aldrich Chemical) and sub-micrometre amorphous B powder (Callery Chemical). Starting materials were lightly mixed in half-gram batches, and pressed into pellets. These pellets were placed on Ta foil, which was in turn placed on Al_2O_3 boats, and fired in a tube furnace under a mixed gas of 95% Ar + 5% H_2 for 1 hour at 600 °C, 1 hour at 800 °C, and 1 hour at 900 °C, and then lightly ground. The resulting powders were pressed into pellets and then fired at 700 °C for 1 hour (sample A) or hot pressed at 10 kbar at 800 °C for 1 hour (sample B). As noted later, this process cannot yet be considered optimum. Disks ~4 mm in diameter and ~1 mm thick were cut from these pellets for property characterization. Magnetization properties were examined in a SQUID magnetometer and a vibrating sample magnetometer (VSM) in applied fields up to 14 T from 4.2 K to above T_c . Figure 1 shows temperature dependences of the zero-field cooled (ZFC) magnetic susceptibility $\chi(T)$ for samples A and B and for commercial MgB_2 powder (99.5%, ~2 μm in diameter) which all exhibit onset T_c values of 37–38 K. Sample A and the commercial powder show smooth transitions with some temperature dependence of χ , while sample B exhibits a shoulder in $\chi(T)$, indicative of non-uniformity in superconducting properties. These transitions are not as sharp as those reported in ref. 5 for MgB_2 prepared in sealed Ta and quartz ampoules, yet we observed more than 90% of a full Meissner effect below 20 K in sample B.

Figure 2 shows results of the VSM examination of the magnetization $M(H)$ over a large range of field and temperature. Magnetic hysteresis characteristic of bulk currents was seen at all temperatures; however, the major part of the hysteresis loops in $M(H)$ closed at fields about half of the upper critical field $H_{c2}(T)$. This is shown in Fig. 2a and Fig. 2a inset. A smaller hysteretic tail in $M(H)$ can also be seen in Fig. 2a, which closes at ~0.7 $H_{c2}(T)$. The value of $H_{c2}(T)$ was determined as the field at which $M(H)$ first deviated from the background, as indicated by the dashed line. Figure 2b shows a plot of $H_{c2}(T)$ and a plot of the Kramer extrapolated field (see below) for sample A. Data for sample B are similar to that shown in the plot, and have been omitted to provide clarity. $H_{c2}(T)$ appears to vary

more slowly than predicted by the Werthamer–Helfand–Hohenberg model⁶, which gives $\mu_0 H_{c2}(0) = 0.7\mu_0 T_c(dH_{c2}/dT)|_{T_c} \approx 14$ T from the rather modest slope of $\mu_0 H_{c2}(T)$ of ~0.5 T K⁻¹ at T_c . In fact $\mu_0 H_{c2}(0)$ appears to be 16–18 T, giving a zero temperature coherence length, $\xi(0) = [\phi_0/2\pi\mu_0 H_{c2}(0)]^{0.5}$, of ~4.3 nm. The value of $\xi(0)$ is thus larger than the typical values (1–2 nm) of high-temperature superconductors (HTSs). These results are consistent with the observations of ref. 7, which were obtained at 30 and 36 K.

The flux-pinning characteristics appear similar to those of Nb_3Sn (ref. 8), for which the parameter $J_c^{0.5}H^{0.25}$ is linear in H over a wide range of temperature (Fig. 3a). This Kramer scaling of the flux-pinning force, $F_p = \mu_0 H J_c \propto H^{1/2}(H_K - H)^2$, is found down to at least 20 K for sample A, as indicated in Fig. 3b. The extrapolated intercept $H_K(T)$ defines an empirical irreversibility line for which J_c tends to zero. The $H_K(T)$ line is directly proportional to $\mu_0 H_{c2}(T)$ (Fig. 2b), a result very different from the $(T_c - T)^{1.5}$ dependence seen for the irreversibility field $H^*(T)$ in HTS materials². The proportionality of $H^*(T)$ and $\mu_0 H_{c2}(T)$ is also similar to that observed for Nb–Ti and Nb_3Sn (ref. 9). However, the curvature of the Kramer extrapolation at the highest fields is also suggestive of a low- J_c superconducting phase that is controlling the global current flow through the entire sample. Even so, using standard Bean formulae¹⁰, we conclude that the magnetization hysteresis width in both SQUID and VSM curves

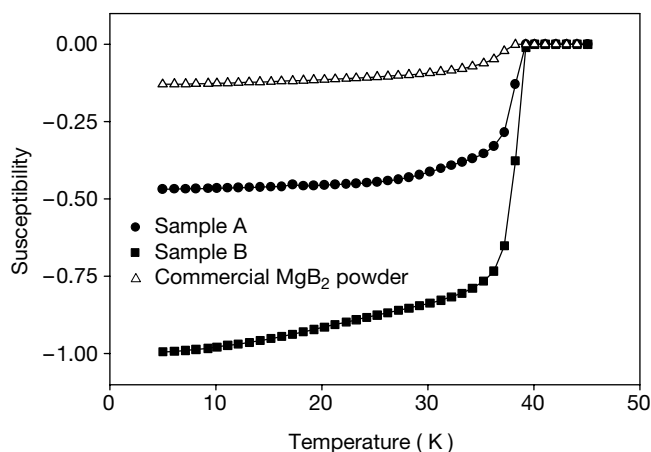


Figure 1 Dimensionless SI magnetic susceptibility as a function of temperature for the different MgB_2 samples studied. These data were obtained from measurements of the magnetic moment after cooling in zero field and warming in a field of 0.5 mT applied parallel to the largest axes of sample A and B. All samples exhibit superconductivity up to 39 K. The broadened transitions of the powder are at least partially due to its small diameter (~2 μm) relative to the superconducting penetration depth. Sample A exhibits about half-full shielding while sample B attains full shielding at low temperatures.

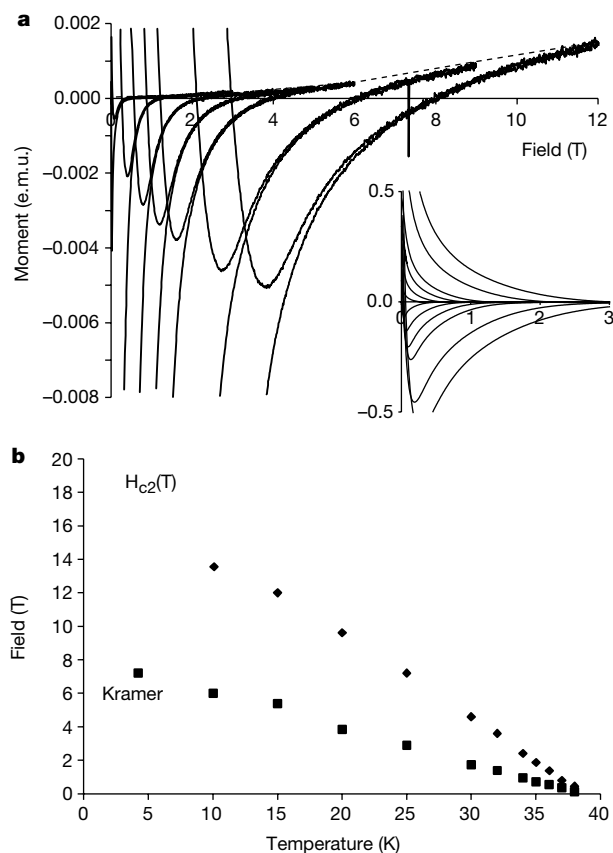


Figure 2 Summary of the field–temperature phase diagram for MgB_2 from magnetization measurements. In **a**, magnetic moment determined by a VSM at 20, 25, 30, 32, 34, 36 and 38 K is plotted, showing a magnified view of the hysteresis loop closure and the background signal (dashed line). The inset shows the overall hysteresis loops at these temperatures. In **b**, the upper critical field H_{c2} (diamonds) and the Kramer extrapolation field (squares) are plotted as a function of temperature. The values of H_{c2} were determined by the field at which the signal intersects the background due to the sample holder and the paramagnetic moment of the sample, and are accurate to ~0.05 T. The Kramer extrapolation field is an estimate of the field of hysteresis loop closure, and represents the loss of flux pinning.

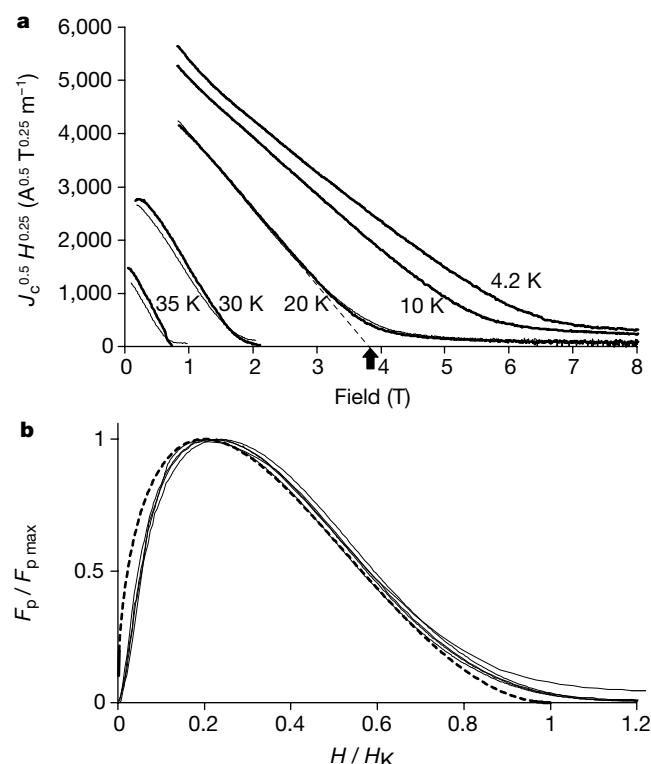


Figure 3 Analysis of flux pinning. In **a**, Kramer plots for samples A (light lines) and B (heavy lines) are derived from the magnetization data from 35 to 4.2 K. The dashed line and the arrow indicate how the extrapolated value H_k was determined at 20 K. In **b**, temperature scaling of pinning force is indicated by the overlap of reduced flux-pinning force curves as a function of reduced field. The data shown (solid lines) are for sample A at 20–37 K, and the heavier dashed line represents the curve shape expected for pinning by grain boundaries⁸.

requires the whole-sample current densities to be $\sim 4 \times 10^4 \text{ A cm}^{-2}$ at 4.2 K, 1 T.

The issue of inhomogeneous and granular behaviour was assessed by correlated magneto-optical (MO), polarized light (PL), and analytical scanning electron microscopy (SEM). Comparison of the PL and MO images in Fig. 4a and b shows a complex multiscale microstructure. Although only $\sim 5\%$ MgO was detected by X-ray diffraction (XRD), the MO images all indicate direct evidence of significant superconducting inhomogeneities. Magnetic flux profiles extracted from the MO images showed type II superconductor behaviour corresponding to both high- and low- J_c regions. The MO images in Fig. 4 reveal high- J_c regions of sizes up to $\sim 150 \mu\text{m}$, in which significant local flux gradients dB/dx were observed at all temperatures from 11 to 38 K, the maximum gradients corresponding to 10^5 A cm^{-2} at 11 K, 0.05 T. These disconnected high- J_c regions are in a low- J_c superconducting matrix, which exhibits almost reversible magnetization. The MO results are consistent with direct transport measurements on sample B, which showed only a rather low, but field-independent, J_c of $\sim 15 \text{ A cm}^{-2}$ at $0 < H < 0.1 \text{ T}$.

Figure 4c and d show that the strongly shielding high- J_c regions are microstructurally subdivided on a scale of $\sim 100 \text{ nm}$. Electron microprobe analysis suggests that the darker central area in Fig. 4c contains a fine mixture of MgB_2 and a boron-rich phase, perhaps due to partial decomposition *in situ*. The XRD patterns of hot-pressed and powder samples were identical, with no indication of texture. Therefore, we conclude that the strongly shielding regions contain a large number of high-angle grain boundaries. Nonetheless, these regions support high current densities of the order of 10^5 A cm^{-2} , which they could not sustain if there was any inherent

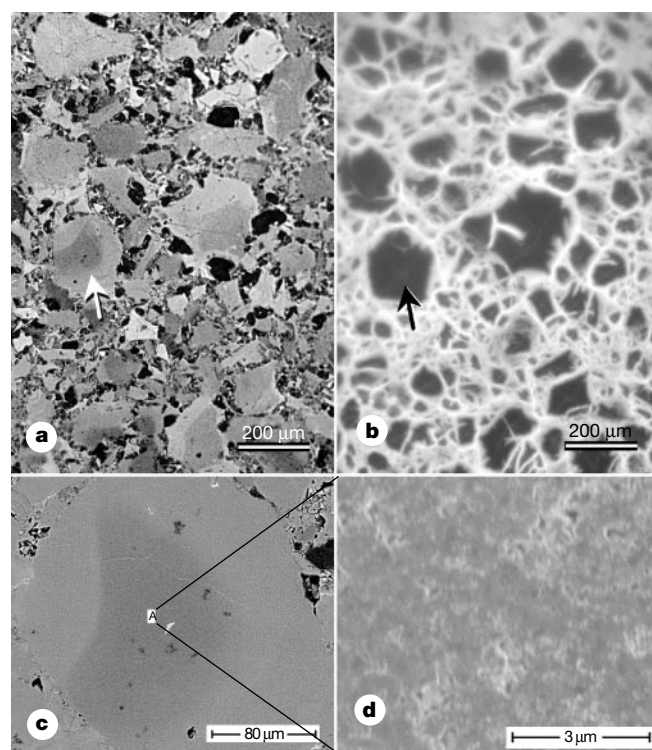


Figure 4 Summary of microstructural and magneto-optical analyses. Polarized light microscope and magneto-optical images of the same area of sample B are compared in **a** and **b**, respectively. Bright regions of **b** indicate areas where magnetic flux has penetrated the sample after a field of 120 mT was applied after cooling the sample in zero field to 11 K. Although the pronounced differences in contrast indicate extensive inhomogeneity of the sample, the dark regions represent areas of strong superconductivity and large screening currents. Image **c** presents a magnified view using SEM backscattered electron imaging of the strongly superconducting region marked with an arrow in **a** and **b**. At higher resolution, image **d**, a secondary electron examination of the central region in **c**, reveals that the area marked by an arrow in **a** and **b** has $\sim 100\text{-nm}$ fine-scale structure.

strong suppression of current across the grain boundaries. Indeed, the true local current densities within the superconducting particles must be significantly larger.

The totality of our data thus leads to the conclusion that MgB_2 is more akin to a low- T_c metallic superconductor than to a high- T_c copper oxide superconductor. A far-from-single-phase sample nevertheless has large current densities that circulate over lengths that are many grain sizes, whether we consider the $\sim 100\text{-nm}$ grains within the strongest $\sim 100\text{-}\mu\text{m}$ agglomerates, or the scale of the whole sample. We have mapped the basic H - T boundary of the superconducting phase, with $H_{c2}(0)$ attaining 16–18 T. This value exceeds that of Nb-Ti but not that of Nb_3Sn , and is rather low for a material with T_c of 39 K. However, the combination of much higher T_c than Nb-base materials and the apparent lack of electromagnetic granularity means that new compounds based on this system could have interesting values of T_c , H_{c2} , and most important of all, J_c , without the need for the high degree of texture that dominates all polycrystalline use of high- T_c materials. □

Received 12 February; accepted 16 February 2001.

1. Nagamatsu, J., Nakagawa, N., Muranaka, T., Zenitani, Y. & Akimitsu, J. Superconductivity at 39 K in magnesium diboride. *Nature* **410**, 63–64 (2001).
2. Yeshurun, Y., Malozemoff, A. P. & Shaulov, A. Magnetic relaxation in high-temperature superconductors. *Rev. Mod. Phys.* **68**, 911–949 (1996).
3. Dimos, D., Chaudhari, P. & Mannhart, J. Superconducting transport properties of grain boundaries in $\text{YBa}_2\text{Cu}_3\text{O}_7$ bicrystals. *Phys. Rev. B* **41**, 4038–4049 (1990).
4. Heinig, N. F., Redwing, R. D., Nordman, J. E. & Larbalestier, D. C. Strong to weak coupling transition in $\text{YBa}_2\text{Cu}_3\text{O}_7$ bicrystals. *Phys. Rev. B* **60**, 1409–1417 (1999).

5. Bud'ko, S. L. *et al.* Boron isotope effect in superconducting MgB_2 . *Phys. Rev. Lett.* **86**, 1877–1880 (2001).
6. Werthamer, N. R., Helfand, E. & Hohenberg, P. C. Temperature and purity dependence of the superconducting critical field, H_{c2} . *Phys. Rev.* **147**, 288–294 (1966).
7. Finnemore, D. K., Ostenson, J. E., Bud'ko, S. L., Lapertot, G. & Canfield, P. C. Thermodynamic and transport properties of superconducting Mg^{10}B_2 . *Phys. Rev. Lett.* (in the press).
8. Dew-Hughes, D. The role of grain boundaries in determining J_c in high-field, high-current superconductors. *Phil. Mag.* **55**, 459–479 (1987).
9. Suenaga, M., Ghosh, A. K., Xu, Y. & Welch, D. O. Irreversibility temperatures of Nb_3Sn and Nb-Ti . *Phys. Rev. Lett.* **66**, 1777–1780 (1991).
10. Campbell, A. M. & Evetts, J. E. Flux vortices and transport currents in type II superconductors. *Adv. Phys.* **21**, 199–428 (1972).

Acknowledgements

Work at the University of Wisconsin was supported by the US Air Force Office of Scientific Research, the US Department of Energy, and the US National Science Foundation through the MRSEC programme; work at Princeton University was supported by the US Department of Energy and the US National Science Foundation.

Correspondence and requests for materials should be addressed to D.C.L. (e-mail: larbales@engr.wisc.edu).

Gate-induced superconductivity in a solution-processed organic polymer film

J. H. Schön*, A. Dodabalapur*, Z. Bao*, Ch. Kloc*, O. Schenker† & B. Batlogg*‡

* Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA
 † University of Konstanz, Department of Physics, D-78457 Konstanz, Germany
 ‡ Solid State Physics Laboratory, ETH Hönggerberg, CH-8093 Zürich, Switzerland

The electrical and optical properties of conjugated polymers have received considerable attention in the context of potentially low-cost replacements for conventional metals and inorganic semiconductors. Charge transport in these organic materials has been characterized in both the doped-metallic and the semiconducting state^{1–4}, but superconductivity has not hitherto been observed in these polymers. Here we report a distinct metal–insulator transition and metallic levels of conductivity in a polymer field-effect transistor. The active material is solution-cast regioregular poly(3-hexylthiophene), which forms relatively well ordered films owing to self-organization, and which yields a high charge carrier mobility ($0.05\text{--}0.1\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$) at room temperature. At temperatures below $\sim 2.35\text{ K}$ with sheet carrier densities exceeding $2.5 \times 10^{14}\text{ cm}^{-2}$, the polythiophene film becomes superconducting. The appearance of superconductivity seems to be closely related to the self-assembly properties of the polymer, as the introduction of additional disorder is found to suppress superconductivity. Our findings therefore demonstrate the feasibility of tuning the electrical properties of conjugated polymers over the largest range possible—from insulating to superconducting.

The mechanism of charge transport in conducting polymers has been the subject of numerous studies since the discovery of metallic conductivity in π -conjugated polymers¹. Such materials can be doped, by adding suitable impurities, to attain conductivity approaching that of copper. A significant feature of such conducting polymers has been that the variation of conductivity with temperature differs from that of metals, such as silver: it does not increase with decreasing temperature over the entire temperature range—this behaviour seems to be related to disorder effects^{5–8}. Polymer films are often composed of two regions, which have to be

distinguished from the viewpoint of carrier transport: relatively ordered regions, and disordered (amorphous) regions⁸. The latter are potential barriers for the electrical transport. Hence, the conduction mechanism in a polymer film comprises a variety of different transport processes, such as intra- and inter-chain transport within the ordered domains and hopping or tunnelling across the amorphous regions. Chemical doping often increases the disorder in such material.

Self-organization of the polymeric chains can lead to remarkably well-ordered films. Regioregular polythiophene has been shown to self-assemble into such films, with the molecules adopting a preferred orientation with respect to the substrate^{9–12}. Consequently, semiconducting regioregular poly(3-alkyl thiophene) can possess a relatively high mobility ($0.05\text{--}0.1\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$) at room temperature^{9,10,13}. This has led to its use in integrated optoelectronics/smart pixels and integrated circuits^{13,14}. Investigations of the morphological, transport and optical properties have shown that the material is nanocrystalline. The high charge-carrier mobility is a result of the relatively low energy barriers to the motion of carriers between adjacent molecules. In comparison, regiorandom polythiophenes possess field-effect mobilities in the range $10^{-5}\text{--}10^{-4}\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$ (ref. 15). Although the properties of conducting and semiconducting polymers have been extensively investigated, superconductivity has not been reported for this class of materials. Organic superconductors known so far include crystals of charge-transfer salts¹⁶, fullerenes¹⁷ and polyacenes¹⁸.

The schematic structure of the field-effect devices based on solution-cast regioregular poly(3-hexylthiophene) films is shown in Fig. 1. The channel resistance has been measured using a four-point probe in order to exclude contact resistance effects. Field-effect mobilities of approximately $0.02\text{--}0.06\text{ cm}^2\text{ V}^{-1}\text{ s}^{-1}$ at room temperature in air can be estimated from the saturation regime of the transistor characteristics. These values are in accordance with reported mobility values for thin films of self-organized regioregular polythiophene^{9,10,13}.

The carrier concentration in the channel region of the polythiophene field-effect device can be adjusted over an extremely wide range by the applied gate bias. In the case of materials held together by van der Waals forces, such as molecular crystals, the absence of defects of the dangling-bond type results in a low density of interface states^{18,19}. Even in nanocrystalline organic semiconductors, the Fermi level is not pinned and can be shifted through the entire density of tail states²⁰. In contrast to chemical doping, gate-induced doping introduces no additional disorder into the material, allowing a systematic study of the electrical properties as a function of the

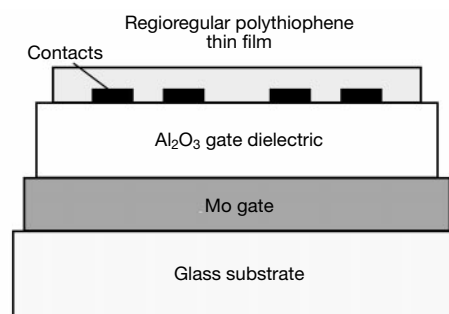


Figure 1 Schematic structure of the polythiophene field-effect device used in this study. Devices were prepared by sputtering a Mo gate electrode ($1\text{ }\mu\text{m}$) onto a glass substrate, followed by the deposition of an Al_2O_3 gate dielectric layer (gate capacitance $C_i = 100\text{--}200\text{ nF cm}^{-2}$). Electrical contacts to the polymer films were prepared by thermal evaporation of gold stripes ($\sim 20\text{ nm}$ thick, and $300\text{--}800\text{ }\mu\text{m}$ long). The distance between contacts varied between 25 and $100\text{ }\mu\text{m}$. Regioregular poly(3-hexylthiophene) films ($\sim 50\text{ nm}$) were prepared by casting from a chloroform solution.

charge carrier density. Moreover, it has been demonstrated that superconductivity can be induced by the field effect^{18,21–23}. The fundamental mechanism underlying gate-induced superconductivity has not yet been definitely determined. Similarities between bulk superconductivity and gate-induced superconductivity in electron-doped C₆₀ (refs 21, 22), as well as tunnelling experiments on pentacene, reveal the importance of electron–phonon interactions. However, the two-dimensional character of the accumulation layer and electron–electron interactions at carrier concentrations exceeding 10¹⁴ cm^{–2} also have to be taken into account.

For hole densities of the order of 10¹² cm^{–2}, polythiophene behaves like a semiconductor. The mobility and the channel resistivity are thermally activated, with an activation energy of approximately 75 meV in the temperature range from 150 to 300 K. Similar behaviour has been reported by Sirringhaus *et al.*¹⁰. Below 100 K, the resistivity becomes more or less temperature independent. This might be ascribed to tunnelling between the individual nanocrystals²⁴. With increasing carrier concentration (gate bias) the temperature dependence of the channel resistance

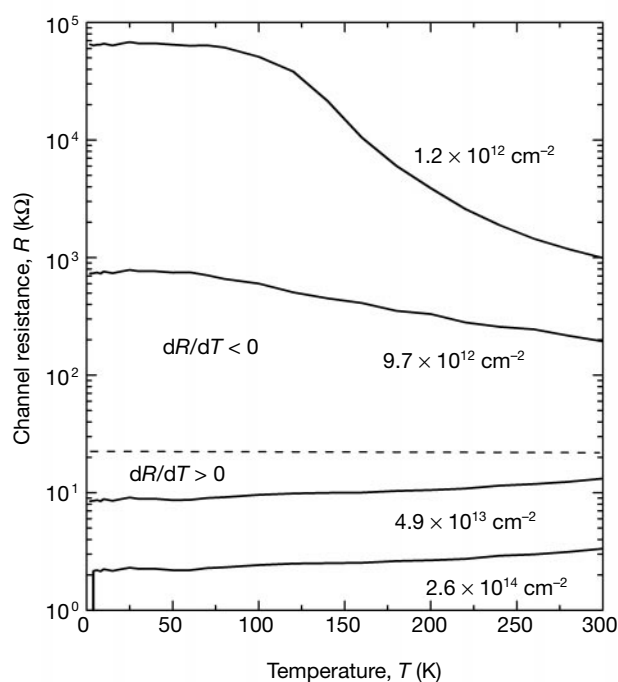


Figure 2 Channel resistance of a polythiophene field-effect device as a function of temperature for various carrier concentrations. The temperature dependence changes from thermally activated (semiconducting) to metallic for hole densities exceeding $4 \times 10^{13} \text{ cm}^{-2}$. Moreover, a transition to a superconducting state below 2.35 K is observed at highest carrier densities. In the absence of concrete evidence of carrier delocalization between chains, we assume that the dimensions of a charge-carrier wavefunction are of the order of the molecular size ($\sim 0.4 \text{ nm}$) along one dimension and the delocalization length along the other. A sheet carrier density of $5 \times 10^{13} \text{ cm}^{-2}$ would correspond to about 1 charge per 12 thiophene rings. A charge density of $2.6 \times 10^{14} \text{ cm}^{-2}$ corresponds to about one charge per 5 thiophene rings. It is easy to visualize why the polythiophene is metallic at such carrier densities, particularly as previous studies have shown that in conjugated polymers the charge carriers are delocalized over many monomer units within a single polymer chain. The conductivity corresponding to a sheet charge density of $5 \times 10^{13} \text{ cm}^{-2}$ is 8 S cm^{-1} , assuming that all the charge is located in the first 1 nm from the interface. The conductivity at temperatures above T_c , at carrier densities sufficient for the onset of superconductivity, is $\sim 80 \text{ S cm}^{-1}$. These values are comparable to the conductivities of other organic superconducting materials. The carrier concentration was inferred from the gate capacitance and the applied gate voltage, as well as from additional Hall effect measurements. The dashed line shows approximately the transition from metallic to insulating behaviour.

changes from thermally activated to metallic; that is, the resistance decreases with decreasing temperature for hole densities exceeding $4 \times 10^{13} \text{ cm}^{-2}$ (see Fig. 2). A similar insulator–metal transition has been observed in chemically doped polythiophene films⁸. We note that the activation energy on the insulating side of the transition decreases with increasing carrier concentration. This might be attributed to electrostatic screening of potential barriers at grain boundaries between the nanocrystallites²⁴. Moreover, the self-organization of the polymer results in similar interchain distances as in oligothiophene single crystals, in which two-dimensional band transport has been observed²⁵.

At hole densities greater than $2.5 \times 10^{14} \text{ cm}^{-2}$, the channel resistance suddenly drops to zero below a critical temperature, T_c of 2.35 K, indicating a transition to a superconducting state. We estimate the critical current density to be about 10^4 A cm^{-2} . The magnetic-field dependence of the phase transition is shown in Fig. 3. Similar gate-induced superconductivity in the two-dimensional channel region has been observed in field-effect devices based on C₆₀ and polyacene^{18,21}. From the shift of T_c , we can estimate the upper critical field H_{c2} for our film as a function of temperature (see Fig. 3 inset). An anisotropic coherence length ξ can be estimated from the extrapolation of H_{c2} to $T = 0 \text{ K}$: this reflects the two-dimensional electrical properties of this self-organized polymer^{9,10}. It is useful to estimate the number of charge carriers induced per unit area in terms of the number of thiophene rings (monomer units) per charge, assuming that the backbone is oriented along the substrate and that the hexyl side chains are approximately normal to the surface. The ability to measure accurately both the carrier density and mobility with the field-effect geometry facilitates the study of the quasi-two-dimensional semiconductor–metal transition. It is easy to visualize why the polythiophene is metallic at very high carrier densities, particularly as previous studies have shown that in conjugated polymers the charge carriers are delocalized over many monomer units within a single polymer chain²⁶.

The room-temperature mobility μ_{RT} is a measure of the order in the polymer film. The superconducting phase transition is shown in

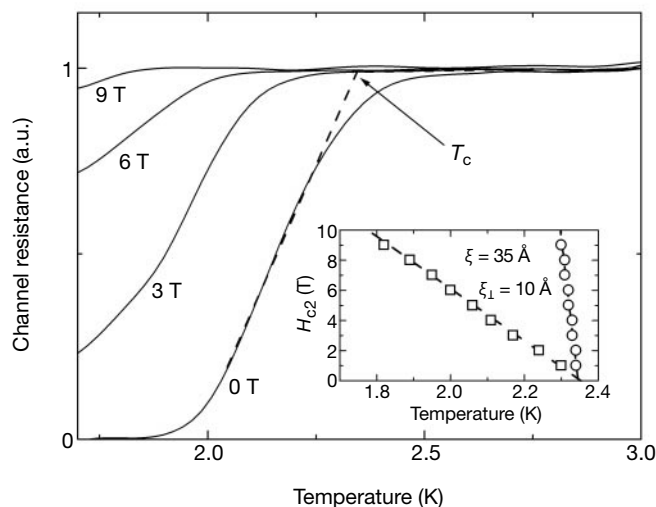


Figure 3 Channel resistance of a polythiophene field-effect device as a function of temperature for various magnetic fields up to 9 T. From the shift of T_c , the upper critical field H_{c2} can be estimated as a function of temperature. Inset, temperature dependence of H_{c2} . Using the Werthamer, Helfand and Hohenberg formula³¹ $H_{c2}(0)$ and the coherence length ξ can be calculated. A coherence length of 35 Å is obtained within the plane of the two-dimensional lamellar structure of the polymer, and a value of approximately 10 Å perpendicular to the plane of the film. This reflects the two-dimensional electrical properties of this self-organized polymer.

Fig. 4 for films with different degrees of disorder. With increasing disorder, T_c shifts to lower temperatures (see Fig. 4 inset) and, furthermore, the transition region becomes broader. In less-ordered material (mobility $< 0.02 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$), the channel resistance remains non-zero down to the lowest experimental temperature of 1.7 K. This may explain why superconductivity has not been observed in chemically doped conjugated polymers. In this case, the disorder introduced by the impurities probably destroys the superconducting phase. However, on the basis of these findings we speculate that bulk superconductivity should be possible in chemically doped, highly ordered organic polymers.

Our results show that the nanocrystalline morphology of regio-regular polythiophene films does not appear to entirely eliminate the superconducting phase. This may be because the film is in fact a network of connected superconducting islands. Superconductivity, due to Josephson coupling of the superconducting islands, has been reported in many such polycrystalline/granular systems²⁷. The shape of the resistance curves and their field dependence are typical of such a granular superconductor. In addition, our observation of a decrease of the critical current density in the more disordered samples (higher normal-state resistance) is in accordance with the picture of a disordered two-dimensional array of Josephson junctions. Global phase coherence is only gradually established upon cooling, and the coupling between the superconducting regions can be suppressed by an external magnetic field. This results in a pronounced resistance decrease, yet zero resistance may not be fully reached. In our study, the current density during the resistance measurements is about $1\text{--}10 \text{ A cm}^{-2}$ assuming a conducting channel of only one or a few monolayers.

We suggest that the granular superconductivity that we observed is closely related to the microscopic texture of the polymer film. Previous studies have revealed the tendency of hexylthiophene-based materials to form small areas ($\sim 50\text{--}150 \text{ \AA}$) that have a high degree of local order^{10,28}. Although the self-organization of the films has been assumed to be facilitated by the hexyl sidechains, it has also been noticed that the thiophene backbone itself has a strong tendency towards π -stacking with significant order^{9–11}. Thus, we suggest that the observed superconductivity originates in patches (islands) that are at least as large as the intrinsic coherence length ξ —which we expect to be of the order of $100 \text{ \AA}$, similar to the measured coherence lengths in the crystalline forms of anthracene,

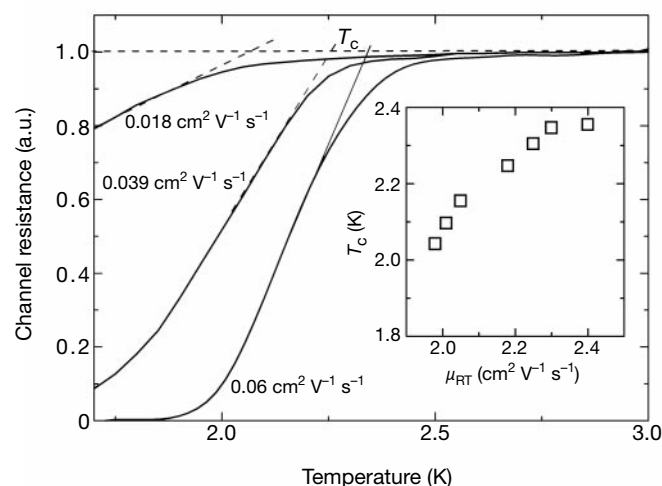


Figure 4 Channel resistance as a function of temperature for polymer thin films with different degrees of disorder. The value of the room-temperature mobility μ_{RT} is taken as a measure of disorder. Inset, the relation between the critical temperature T_c and μ_{RT} (that is, disorder). The transition region broadens with increasing disorder. The current density during the measurement was $\sim 1 \text{ A cm}^{-2}$.

tetracene, or pentacene¹⁸, other examples of van-der-Waals-bonded conjugated organic materials. This estimated lower limit to the size of the superconducting islands agrees with the above-mentioned structural studies⁹. Our estimation of the 'upper critical field', as defined in Fig. 3, is not necessarily determined only by the intrinsic coherence length, but is also influenced by the field-induced reduction of the coupling between the superconducting grains. The resulting numerical values of an effective coherence length ξ of approximately $35 \text{ \AA}$ are consistent with such an estimate. We are at present characterizing the dependence of the superconductivity on sample morphology and magnetic field, and expect that our results will shed more light on this matter.

The present results also indicate that superconductivity may be found in other partially ordered conjugated organic materials. A number of such materials have been shown²⁴ to possess room-temperature mobilities exceeding $0.1 \text{ cm}^2 \text{ V}^{-1} \text{ s}^{-1}$. The ease of processing of these organic/polymeric materials, together with their demonstrated ability to be fabricated into sophisticated circuitry^{13,14,29}, suggests possible future applications in the field of superconducting electronics. We also note that superconductivity-based phenomena are one of the approaches pursued for quantum information processing³⁰. □

Received 14 November 2000; accepted 1 January 2001.

- Chiang, C. K. *et al.* Electrical conductivity in doped polyacetylene. *Phys. Rev. Lett.* **39**, 1098–1101 (1977).
- Skotheim, T. A., Elsenbaumer, R. L. & Reynolds, J. R. (eds) *Handbook of Conjugated Polymers* (Marcel Dekker, New York, 1996).
- Mark, J. (ed.) *Physical Properties of Polymers Handbook* (AIP Press, New York, 1996).
- Nalwa, H. S. *Handbook of Organic Conductive Molecules and Polymers* (Wiley, Chichester, 1997).
- Aleshin, A., Kiebooms, R., Reghu, M. & Heeger, A. J. Electronic transport in doped poly(3,4-ethylenedioxythiophene) near the metal-insulator transition. *Synth. Met.* **90**, 61–68 (1997).
- Ahlsgog, M. & Menon, R. Charge transport in conducting polymers near the critical and insulating regimes. *Phys. Status Solidi B* **205**, 305–310 (1998).
- Kim, J. H. *et al.* Structural aspects of metal-insulator transition in doped conducting polymers. *Synth. Met.* **84**, 71–72 (1997).
- Masubuchi, S. & Kazama, S. Intrinsic transport properties in electrochemically prepared polythiophene doped with PF_6^- . *Synth. Met.* **74**, 151–158 (1995).
- Bao, Z., Dodabalapur, A. & Lovinger, A. J. Soluble and processable regioregular poly(3-hexylthiophene) for thin film field-effect transistor applications with high mobility. *Appl. Phys. Lett.* **69**, 4108–4110 (1996).
- Sirringhaus, H. *et al.* Two-dimensional charge transport in self-organized, high-mobility conjugated polymers. *Nature* **401**, 685–688 (1999).
- Österbacka, R., An, C. P., Jiang, X. M. & Vardeny, Z. V. Two-dimensional electronic excitations in self-assembled conjugated polymer nanocrystals. *Science* **287**, 839–842 (2000).
- McCulloch, R. D., Tristram-Nagle, S., Williams, S. P., Lowe, R. D. & Jayaraman, M. Self-orienting head-to-tail poly(3-alkylthiophenes)—New insights on structure-property relationship in conducting polymers. *J. Am. Chem. Soc.* **115**, 4910–4911 (1993).
- Sirringhaus, H., Tessler, N. & Friend, R. H. Integrated optoelectronic devices based on conjugated polymers. *Science* **280**, 1741–1744 (1998).
- Dodabalapur, A. *et al.* Organic smart pixels. *Appl. Phys. Lett.* **73**, 142–144 (1998).
- Kobashi, M. & Takeuchi, H. Inhomogeneity of spin-coated and cast non-regioregular poly(3-hexylthiophene) films. Structures and electrical and photophysical properties. *Macromolecules* **31**, 7273–7278 (1998).
- Ishiguro, T., Yamaji, K. & Saito, G. *Organic Superconductors* (Springer, Berlin, 1998).
- Gunnarsson, O. Superconductivity in fullerides. *Rev. Mod. Phys.* **69**, 575–606 (1997).
- Schön, J. H., Kloc, Ch. & Batlogg, B. Superconductivity in molecular crystals induced by charge injection. *Nature* **406**, 702–704 (2000).
- Schön, J. H., Berg, S., Kloc, Ch. & Batlogg, B. Ambipolar pentacene field-effect transistors and inverters. *Science* **287**, 1022–1023 (2000).
- Schön, J. H., Kloc, Ch. & Batlogg, B. Normal and inverse Meyer-Neldel rule in nanocrystalline pentacene field-effect transistors. *Phys. Status Solidi B* **121**, R4–R5 (2000).
- Schön, J. H., Kloc, Ch., Haddon, R. C. & Batlogg, B. A Superconducting field-effect switch. *Science* **288**, 656–658 (2000).
- Schön, J. H., Kloc, Ch. & Batlogg, B. Superconductivity at 52 K in hole-doped C_{60} . *Nature* **408**, 549–552 (2000).
- Ahn, C. H. *et al.* Electrostatic modulation of superconductivity in ultrathin $\text{GdBa}_2\text{Cu}_3\text{O}_{7-x}$ films. *Science* **284**, 1152–1155 (1999).
- Horowitz, G. & Hajlaoui, M. E. Mobility in polycrystalline oligothiophene field effect transistors dependent on grain size. *Adv. Mater.* **12**, 1046–1050 (2000).
- Schön, J. H., Kloc, Ch. & Batlogg, B. Charge transport in oligothiophene single crystals. *Synth. Met.* **115**, 75–78 (2000).
- Knupfer, M. *et al.* Size of electron-hole pairs in π -conjugated systems. *Phys. Rev. Lett.* **83**, 1443–1446 (1999).
- Simanek, E. *Inhomogeneous Superconductors: Granular and Quantum Effects* (Oxford Univ. Press, New York, 1994).
- Lovinger, A. J., Katz, H. E. & Dodabalapur, A. Direct imaging of conducting and insulating submolecularly wide pathways in an organic semiconductor. *Chem. Mater.* **10**, 3275–3278 (1998).

29. Crone, B. *et al.* Large-scale complementary integrated circuits based on organic transistors. *Nature* **403**, 521–523 (2000).
30. Mooij, J. E. *et al.* Josephson persistent-current qubit. *Science* **285**, 1036–1039 (1999).
31. Verthamer, N. R., Helfand, E. & Hohenberg, P. C. Temperature and purity dependence of the superconducting field, H_{c2} , III. Electron spin and spin-orbit effects. *Phys. Rev.* **147**, 295–303 (1966).

Acknowledgements

We thank E. A. Chandross, B. Crone, H. E. Katz, H. Y. Hwang, A. J. Lovinger and T. Siegrist for discussions, and E. Bucher for the use of equipment.

Correspondence and requests for materials should be addressed to J.H.S. (e-mail: hendrik@lucent.com).

An efficient room-temperature silicon-based light-emitting diode

Wai Lek Ng*, M. A. Lourenço*, R. M. Gwilliam*, S. Ledain†, G. Shao† & K. P. Homewood*

* School of Electronic Engineering, Information Technology & Mathematics;

† School of Mechanical and Materials Engineering, University of Surrey, Guildford, Surrey, GU2 7XH, UK

There is an urgent requirement for an optical emitter that is compatible with standard, silicon-based ultra-large-scale integration (ULSI) technology¹. Bulk silicon has an indirect energy bandgap and is therefore highly inefficient as a light source, necessitating the use of other materials for the optical emitters. However, the introduction of these materials is usually incompatible with the strict processing requirements of existing ULSI technologies. Moreover, as the length scale of the devices decreases, electrons will spend increasingly more of their time in the connections between components; this interconnectivity problem could restrict further increases in computer chip processing power and speed in as little as five years. Many efforts have therefore been directed, with varying degrees of success, to engineering silicon-based materials that are efficient light emitters^{2–7}. Here, we describe the fabrication, using standard silicon processing techniques, of a silicon light-emitting diode (LED) that operates efficiently at room temperature. Boron is implanted into silicon both as a dopant to form a p–n junction, as well as a means of introducing dislocation loops. The dislocation

loops introduce a local strain field, which modifies the band structure and provides spatial confinement of the charge carriers. It is this spatial confinement which allows room-temperature electroluminescence at the band-edge. This device strategy is highly compatible with ULSI technology, as boron ion implantation is already used as a standard method for the fabrication of silicon devices.

There have been great efforts over the past decade to obtain useful, that is, technologically viable and efficient, light emission from silicon both in the visible and infrared regions of the spectrum. In the visible regions, porous silicon² and other quantized systems, such as silicon/silicon dioxide superlattices³ and silicon nanoprecipitates in silicon dioxide⁴, have been the main emphasis. In the infrared region, systems such as erbium in silicon⁵, silicon/germanium⁶ and, more recently, iron disilicide⁷ offer potential routes. No approach has so far been applied commercially. The reasons for this are a combination of the lack of genuine or perceived compatibility with conventional ULSI technology and, in the case of infrared emitters, high thermal quenching giving very poor room-temperature efficiencies.

Here we use a new approach—dislocation engineering, using conventional ULSI technology—that gives efficient light emission in silicon at room temperature.

Because of its indirect bandgap, silicon is fundamentally a poor emitter of light. The main reason for this is that fast non-radiative recombination routes dominate the slower radiative route in this material. Indeed, in bulk silicon, at room temperature, radiative emission is normally entirely absent. However, if recombination through the non-radiative routes can be prevented, the radiative emission could in principle be enhanced. Non-radiative recombination is the result of diffusion of carriers to point defects in the silicon where efficient non-radiative recombination then occurs. Despite the low defect concentrations in good quality silicon this non-radiative route is always completely dominant. A way of enhancing the radiative efficiency would be to prevent the carrier diffusion. If silicon can be formed as clusters then strong band-edge photoluminescence is possible⁸, but thus far this has only been achieved by incorporating these clusters inside large bandgap insulating oxides. However, the insulating matrix prevents efficient carrier injection, making devices difficult to produce. Similar but much weaker (with efficiencies up to about 8×10^{-6}) band-edge emission has also been observed in laser-recrystallized silicon⁹, but no explanation of its origin was presented.

We have made use of the controlled introduction of dislocation loops using conventional ion implantation and thermal processing. The dislocation loop array, if appropriately produced, introduces a strain field in three dimensions that modifies the bandgap of the

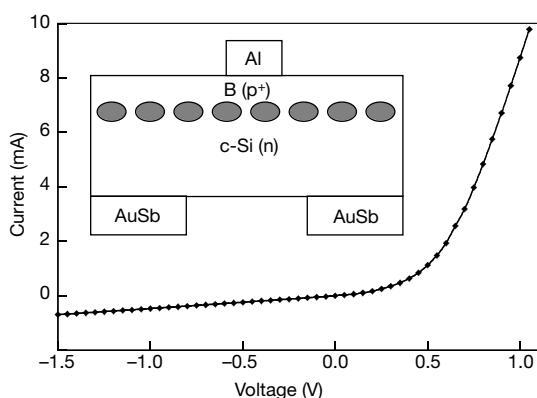


Figure 1 The current–voltage plot for the device measured at room temperature. Inset, a schematic of the light-emitting diode (LED) device. The top and bottom ohmic contacts are formed by Al and AuSb respectively. The infrared light is emitted through the window left in the bottom contact.

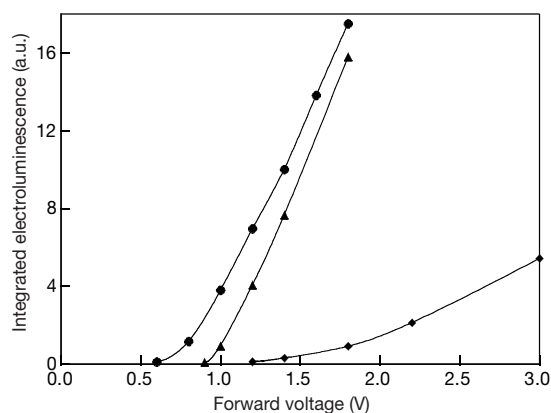


Figure 2 Plots of the integrated electroluminescence intensity as a function of applied forward voltage at various temperatures: 80 K (diamonds), 180 K (triangles) and 300 K (circles).

silicon in such a way that the silicon itself can be used to provide spatial confinement in three dimensions. To minimize the number of process steps, in the device described, boron implantation has been used both to introduce the dislocation loop array and as the p-type dopant to form a p–n junction in an n-type silicon substrate. However, another implant species such as the host silicon could be used to form the dislocations so that the dislocation engineering and subsequent doping to form the p–n junction can be achieved independently. The device operates as a conventional light-emitting diode under forward bias.

A simple diagram of the device is shown in the inset to Fig. 1. The device reported here was made by implanting boron at a dose of $1 \times 10^{15} \text{ cm}^{-2}$ at an energy of 30 keV. The sample was subsequently annealed in a nitrogen atmosphere for 20 minutes at 1,000 °C to form the dislocation loop array and activate the boron dopants. The implants were made into a device grade CZ (Czochralski) n-type silicon substrate of resistivity 2–4 $\Omega \text{ cm}$. An important point to emphasize at this stage is that all these process steps are entirely conventional and are therefore completely compatible with ULSI technology, allowing immediate implementation on a standard fabrication line. The array of dislocation loops formed has been observed using cross-sectional transmission electron microscopy. The array is situated in a planar region parallel to, and around 100 nm from, the p–n junction. The dislocation loops are typically about 80–100 nm in diameter and are spaced around 20 nm apart. The strain field at the edge of the dislocation loop is high and falls off around each loop approximately with the inverse of distance. The magnitude and sign of the stress at the edge of the dislocation loop which determines the confining potential can be calculated using the standard elastic theory of dislocations¹⁰ and the known values of the Poisson's ratio, 0.42, and Young's modulus, 113 GPa, for silicon. We obtain a value for the maximum stress at the outside edge (1.2 Å out) of the interstitial dislocation loop of 25–50 GPa, which will cause an increase in bandgap energy at the edge of 325–750 meV. A complete description of the stress field across the device, which is just the superposition of the stress field of all the dislocations, is complicated and will vary in detail across and between samples. However, given the inverse distance dependence of the stress around the loops, the local stress at any point is entirely dominated by the superposition of the stress field of only the closest dislocations. As carriers are injected across the junction they encounter a blocking potential that varies in height depending on the direction of their instantaneous velocity, from about 750 meV in the direction of and across a loop, down to about 20 meV at a point equidistant between two adjacent dislocations. It is this blocking potential that confines carriers close to the junction region. The varying potential falls away approximately as inverse distance back towards the junction at

which point it is less than about 10 meV. Consequently, the barrier height and the onset of injection is not significantly changed. Any carriers moving back across the junction are, under steady state conditions, dynamically replaced by others moving in the forward direction.

After implantation and annealing, ohmic contacts consisting of AuSb eutectic and Al were applied to the back and front of the sample respectively. The ohmic contacts were sintered at 360 °C for 2 min. The top contact was 1 mm in diameter and, as the silicon substrate is still nearly transparent at the emission wavelength of 1,150 nm, a window was left in the larger back contact to allow the electroluminescence (EL) through. Front window devices have also been fabricated. Current–voltage (*I*–*V*) measurements were made between the back and front contacts to check that the device is behaving as a diode. An *I*–*V* plot, characteristic of the fully processed and working device, is shown in Fig. 1.

The device was mounted into a mechanical holder inside a continuous flow liquid-nitrogen cryostat. Light from the device was focused into a conventional 0.5 m spectrometer and collected by a liquid-nitrogen-cooled germanium p–i–n detector. EL measurements were then taken from 80 K to above room temperature. The onset of EL was observed as the diode turns on under forward bias. No EL is observed under reverse bias. The integrated EL intensity as a function of applied forward voltage is shown in Fig. 2 for several temperatures. The turn-on voltage at room temperature is typical of a conventional silicon diode operating under forward injection. The full EL spectra taken, as a function of temperature, at a forward current of 50 mA, are shown in Fig. 3. The low-temperature EL spectrum from the device shows the main features expected from emission at the silicon band edge, including, at 1,190 nm, the phonon replica of the main peak at 1,130 nm. The room-temperature EL spectrum has the main peak at 1,160 nm with a full-width at half-maximum of 80 nm. No EL or photoluminescence (PL) is observed at any other wavelengths, even at low temperatures—a result that we again attributed to the close spatial confinement of carriers within regions of bulk silicon. The temperature dependence of the integrated EL intensity is shown in Fig. 4. In most systems, demonstrated for light emission in silicon^{5,6,7} the EL quenches very strongly with increasing temperature, making practical room-temperature devices problematic. Here, it can be seen that the EL intensity actually increases as we go up in temperature. A similar behaviour has been observed but not explained in laser-recrystallized silicon⁹. The device is therefore, as required, most efficient at room temperature and above. Inset in Fig. 4 is a plot of the integrated PL intensity as a function of measurement tempera-

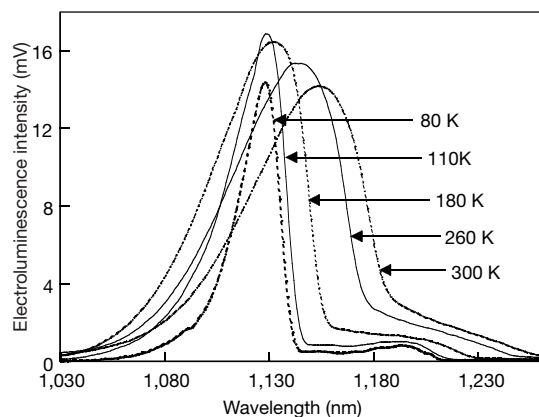


Figure 3 Spectra of the electroluminescence intensity against wavelength at various temperatures. The device was operated at a forward current of 50 mA for all temperatures.

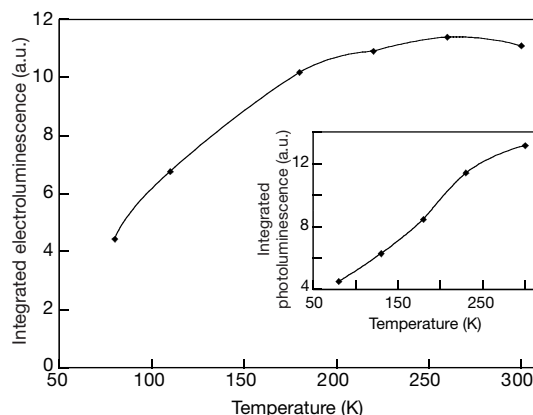


Figure 4 A plot of the integrated electroluminescence intensity as a function of measurement temperature. Inset, a plot of the integrated photoluminescence intensity as a function of temperature. The solid lines are provided as a guide to the eye. The photoluminescence was excited by the 488 nm line of an argon laser at a power of 150 mW.

ture. The integrated PL also increases with temperature in the same manner as the EL, indicating that the temperature dependence is intrinsic to the recombination, rather than the injection mechanism. The strong temperature quenching of PL and EL in most semiconductors is primarily the result of the strong temperature dependence of the competing non-radiative routes—the band-to-band transition is relatively temperature-independent. In our case spatial localization of the radiative carrier population decouples it from any non-radiative recombination occurring elsewhere, thus eliminating the luminescence quenching. A full understanding of the form of the weak increase of the integrated intensity with temperature seen requires further investigation; it could be associated with increased scattering within the confined carrier populations and the increase in the effective density of band states with temperature. We have also carried out EL frequency-resolved measurements that give a device response time, at room temperature, of $18 \pm 2 \mu\text{s}$. The PL and EL are both superlinear with excitation power and drive current respectively, meaning that the device improves further as we drive the device harder. In ref. 9 a similar dependence of the luminescence intensity is attributed, owing to a shift in emission wavelength, to sample heating. However, in our case, no such shift is observed.

The external quantum efficiency of our unpackaged planar device has been measured, based only on the light emitted just through the back window. At a 100 mA forward current the emitted light is $19.8 \mu\text{W}$ giving an external quantum efficiency of $(2.0 \pm 0.1) \times 10^{-4}$ at room temperature. We also obtain significant edge emission from the device, of a further $80 \mu\text{W}$, but this is not included in our estimate of the quantum efficiency above. Taking it into account, our device has a quantum efficiency of 10^{-3} . The emitted power from the face of the device was measured by placing it immediately adjacent to a large-area calibrated power meter. The power meter is an Ophir Laser Star fitted with a PD300 IR head; the instrument is certificated and calibrated with standards traceable to the National Institute of Standards. Our front emitting devices have the same efficiencies as the back emitting devices. For comparison, commercially available GaAs infrared LEDs have typical external efficiencies of 10^{-2} . However, commercial packing of LEDs, in particular by minimizing the considerable internal reflection losses of the planar device and by collecting the edge emission, can improve the external efficiency by up to a factor of about 15. Our first, non-optimized devices are therefore already within a factor of 3 or so of the efficiencies achieved in conventional optimized LED devices.

We believe that the device demonstrated is the most likely candidate currently for implementing efficient light sources in silicon. Such a device would also form the basis for the development of an injection laser based on the same principles but with the incorporation of an optical cavity. The approach itself is not limited to silicon but could be applied to other materials, particularly other indirect materials and silicon alloys. For example, going from germanium to silicon to silicon carbide could produce devices, using the same approach, that could span the near-infrared region, including the $1.3 \mu\text{m}$ and $1.5 \mu\text{m}$ wavelength regions of the spectrum, important for optical fibre communications, and up to the ultraviolet region. □

Received 23 October 2000; accepted 29 January 2001.

1. European Commission Technology Roadmap—Optoelectronic Interconnects for Integrated Circuits (eds Forchel, A. & Malinverni, P.) (Office for Official Publications of the European Communities, Luxembourg, 1998).
2. Hirschman, K. D., Tybesskov, L., Duttgupta, S. P. & Fauchet, P. M. Silicon-based visible light-emitting devices integrated into microelectronic circuits. *Nature* **384**, 338–341 (1996).
3. Lu, Z. H., Lockwood, D. J. & Baribeau, J.-M. Quantum confinement and light emission in SiO_2/Si superlattices. *Nature* **378**, 258–260 (1995).
4. Komoda, T. et al. Visible photoluminescence at room temperature from microcrystalline silicon precipitates in SiO_2 formed by ion implantation. *Nucl. Instrum. Methods B* **96**, 387–391 (1995).
5. Zheng, B. et al. Room-temperature sharp line electroluminescence at $\lambda = 1.54 \mu\text{m}$ from an erbium-doped, silicon light-emitting diode. *Appl. Phys. Lett.* **64**, 2842–2844 (1994).

6. Vescan, L. & Stoica, T. Room-temperature SiGe light-emitting diodes. *J. Luminescence* **80**, 485–489 (1999).
7. Leong, D., Harry, M., Reeson, K. J. & Homewood, K. P. A silicon/iron disilicide light-emitting diode operating at a wavelength of $1.5 \mu\text{m}$. *Nature* **387**, 686–688 (1997).
8. Tybesskov, L., Moore, K. L., Hall, D. G. & Fauchet, P. M. Intrinsic band-edge photoluminescence from silicon clusters at room temperature. *Phys. Rev. B* **54**, R8361–R8364 (1996).
9. Sveinbjörnsson, E. O. & Weber, J. Room temperature electroluminescence from dislocation rich silicon. *Appl. Phys. Lett.* **69**, 2686–2688 (1996).
10. Hirth, J. P. & Lothe, J. *Theory of Dislocations* 2nd edn, 63 (John Wiley & Sons, New York, 1982).

Correspondence and requests for materials should be addressed to K.P.H. (e-mail: k.homewood@eim.surrey.ac.uk).

Dating of the oldest continental sediments from the Himalayan foreland basin

Yani Najman^{*§}, Malcolm Pringle[†], Laurent Godin^{‡§} & Grahame Oliver[¶]

^{*} Fold and Fault Research Project, Department of Geology and Geophysics, University of Calgary, 2500 University Drive NW, Calgary, AB, Canada T2N 1N4

[†] Scottish Universities Environmental Research Centre, Scottish Enterprise Technology Park, Rankine Avenue, East Kilbride G75 0QF, UK

[‡] Department of Earth Sciences, Oxford University, Parks Road, Oxford OX1 3PR, UK

[¶] Crustal Geodynamics Group, School of Geography and Geosciences, Purdie Building, The University, St Andrews, Fife KY16 9ST, UK

A detailed knowledge of Himalayan development is important for our wider understanding of several global processes, ranging from models of plateau uplift to changes in oceanic chemistry and climate^{1–4}. Continental sediments 55 Myr old found in a foreland basin in Pakistan⁵ are, by more than 20 Myr, the oldest deposits thought to have been eroded from the Himalayan metamorphic mountain belt. This constraint on when erosion began has influenced models of the timing and diachrony of the India–Eurasia collision^{6–8}, timing and mechanisms of exhumation^{9,10} and uplift¹¹, as well as our general understanding of foreland basin dynamics¹². But the depositional age of these basin sediments was based on biostratigraphy from four intercalated marl units⁵. Here we present dates of 257 detrital grains of white mica from this succession, using the ⁴⁰Ar–³⁹Ar method, and find that the largest concentration of ages are at 36–40 Myr. These dates are incompatible with the biostratigraphy unless the mineral ages have been reset, a possibility that we reject on the basis of a number of lines of evidence. A more detailed mapping of this formation suggests that the marl units are structurally intercalated with the continental sediments and accordingly that biostratigraphy cannot be used to date the clastic succession. The oldest continental foreland basin sediments containing metamorphic detritus eroded from the Himalaya orogeny therefore seem to be at least 15–20 Myr younger than previously believed, and models based on the older age must be re-evaluated.

The Balakot Formation, located in the Hazara–Kashmir Syntaxis of Northern Pakistan, is a continental foreland basin sedimentary sequence that contains detritus eroded from the India–Eurasia suture zone and the metamorphic rocks of the Himalaya¹¹. The Balakot Formation overlies the Palaeocene shallow marine Patala Formation (Figs 1 and 2) and consists of a >8-km-thick fossil-free

§ Present addresses: Department of Geology & Geophysics, Edinburgh University, West Mains Road, Edinburgh EH9 3JW, UK (Y.N.); Department of Earth Sciences, Simon Fraser University, 8888 University Drive, Burnaby BC, V5A 1S6, Canada (L.G.).

clastic red bed succession, within which are intercalated four discrete grey fossiliferous marl bands (Fig. 2). Nummulites and assilines biostratigraphy of the marl bands has constrained the succession to 55–50 Myr (latest Palaeocene to Mid Eocene)⁵. On this basis, the Balakot Formation has been interpreted as being by far the oldest continental sedimentary succession in the foreland basin that contains detritus from the Himalayan metamorphic belt; the start of similar sedimentation did not occur elsewhere in the basin until more than 20 Myr later (Fig. 1): from Bangladesh⁸, through India¹³ to Pakistan, although in the Kohat region of Pakistan there was a brief interlude of continental sedimentation, the 100–150-m-thick Mami Khel deposits, in the late Early Eocene (~50 Myr) before marine conditions resumed¹⁴.

Our new data show that the Balakot Formation is at least 15–20 Myr younger than previously believed. ⁴⁰Ar–³⁹Ar ages of 257 detrital single white micas separated from nine samples representing the entire thickness of the Balakot Formation show a significant population of micas 36–40 Myr old (Fig. 3a, and Supplementary Information, Table 1). Because a detrital mineral age cannot be younger than its host sediment depositional age, it is clear that the mica ages and biostratigraphic ages of 55–50 Myr deduced from the fossiliferous marl units are incompatible. One explanation for this apparent discrepancy could be that the mica ages do not represent cooling ages characteristic of the original source area but instead have been reset by metamorphism after deposition in the foreland basin and/or low-temperature alteration. We reject this contention, at least for the main 36–40-Myr mode, for the following reasons. First, all of the micas from the young population analysed so far by step-heating techniques (eight in total) show flat plateaux indicating no significant low-temperature alteration (Fig. 3b). (We do note, as expected, that some of the micas >200 Myr old do show some evidence of alteration; the plateaux are not shown. In addition,

it is possible that at least some of the apparently youngest grains analysed by total-fusion techniques, that is, the four grains <32 Myr old in Fig. 3a, have been altered.) Second, electron microprobe traverses of the micas (Supplementary Information) show that ~50% of the micas analysed show no evidence of alteration (as determined by alkali loss) and that, of the remaining 50%, alkali loss is largely confined to grain edges; in most of these latter grains the loss is insubstantial. Last, epizone/greenschist facies conditions (above ~375 °C) are required to reset Ar–Ar age patterns in detrital illite–white mica¹⁵. Petrographic analysis shows that the Balakot Formation has been subjected to only relatively low grades of burial

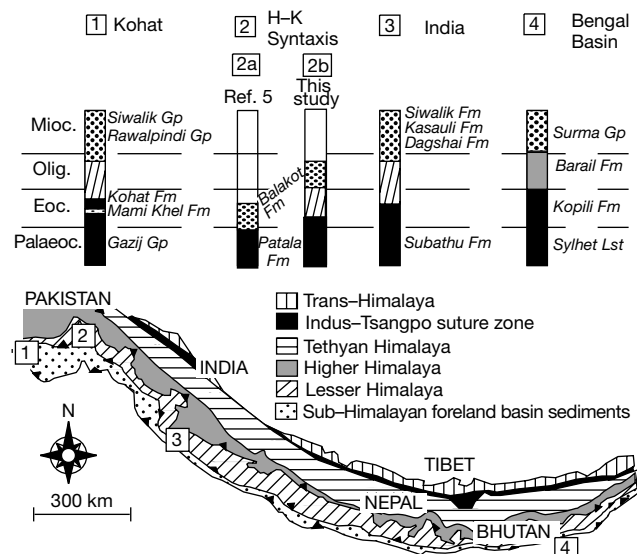


Figure 1 Geological map of the Himalaya showing the location of the study area (2) and a comparison of the foreland basin stratigraphy along strike. Stratigraphic panels are shown from Kohat Region, Pakistan¹⁴ (location and stratigraphic summary (1)), through the Hazara–Kashmir study area (2) (2a, previously accepted stratigraphy, after Bossart and Ottiger⁵; 2b, revised stratigraphy based on this study), to Himachal Pradesh, NW India¹³ (3) and the Bengal Basin, Bangladesh⁸ (4). Legend for the foreland basin stratigraphic columns: black, shallow marine sediments; grey, deltaic Indian craton-derived sediments; stippled, continental orogen-derived sediments; hatched, no sediments preserved. Note that, with the previously accepted stratigraphy, the Balakot Formation of the Hazara–Kashmir Syntaxis, Pakistan is significantly older than the earliest continental deposits anywhere along strike in the basin.

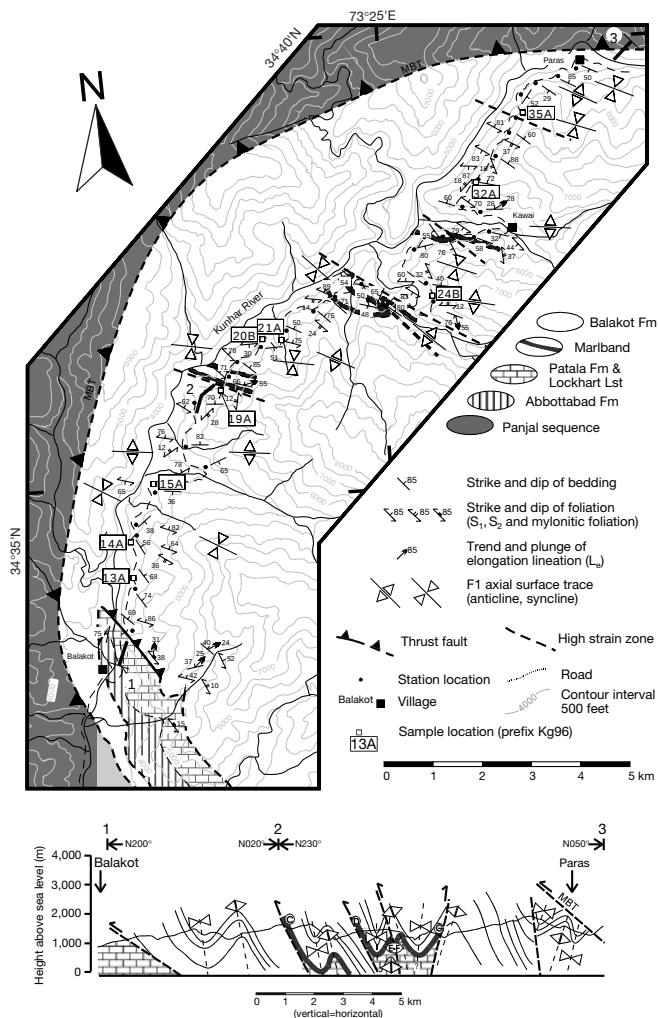


Figure 2 Geological map and cross-section of the Balakot Formation, Kaghan Valley, Pakistan. These new data contrast with mapping undertaken by Bossart and Ottiger⁵ (compare with their Fig. 8), who interpreted the Balakot Formation as a steeply north-dipping normal stratigraphic succession, with four intercalated fossiliferous marl bands (labelled C–G). Here we show the presence of south dipping strata, folds, faults and high strain zones. The fossiliferous marl bands (labelled C–G, the same annotations as those used by Bossart and Ottiger⁵ in their Fig. 8, for easy comparison) outcrop in the core of anticlines or at high strain zones. We therefore interpret the marl bands as part of an underlying formation, now structurally intercalated with the Balakot Formation and exposed by subsequent deformation. Note also that the Patala–Balakot Formation boundary, interpreted by Bossart and Ottiger⁵ as a conformable contact, is here interpreted as a thrust contact, on the basis of its tectonized appearance and shear sense indicators. The cross-section was drawn from point 1, through 2 to 3, as shown on the map. The section is located west of marl band sites E and F. Because the structures plunge moderately to the northwest, marl band sites E and F are located below the surface.

metamorphism (prehnite–pumpellyite facies, that is, corresponding to anchizone, $\sim 200\text{--}375^\circ\text{C}$)^{5,16}. This is supported by illite crystallinity measurements, performed on the $<2\text{-}\mu\text{m}$ predominantly diagenetic component of the rock, which give an average Hb_{rel} value of 254 ± 43 (Supplementary Information, Table 3). These values correspond to the lower anchizone and upper diagenetic zones, implying that these rocks have been subjected to metamorphic conditions of $\sim 200 \pm 50^\circ\text{C}$ ^{17–19}, which is insufficient to reset $^{40}\text{Ar}\text{--}^{39}\text{Ar}$ ages in detrital white mica.

We are therefore confident that the $^{40}\text{Ar}\text{--}^{39}\text{Ar}$ single-crystal mica dates are detrital mineral ages representing the timing of cooling and exhumation in the Himalayan source region, before erosion and deposition with the Balakot Formation, and therefore serve as a robust maximum estimate for the age of sedimentation.

Bossart and Ottiger⁵ mapped the Balakot Formation in the Kaghan valley as a steeply north-dipping, homoclinal, normal stratigraphic succession. However, our detailed mapping (Fig. 2), aided by substantial new exposure owing to extensive road building, revealed that the Balakot Formation is intensely folded, with the fossiliferous marl bands outcropping near the cores of anticlines or high-strain zones. These zones are characterized by a penetrative cleavage nearly parallel to bedding, suggestive of intense bedding transposition. Systematic near-vertical slicken-sides and elongated reduction spots indicate cleavage/parallel slip in the vertical direction. We therefore argue that the marl units are part of an underlying formation, exposed by subsequent deformation related to shear folding and thrusting (Fig. 2b). In addition, we disagree with the interpretation of Bossart and Ottiger⁵ that the contact between the Balakot Formation and underlying Palaeocene Patala Formation is conformable. In the region described by Bossart and Ottiger (Behrin Katha), there is insufficient exposure for adequate interpretation. North of Balakot, the Patala Formation outcrops in a highly tectonized zone, imbricated between the west-dipping Abbottabad Formation to the south, and the north-east-dipping Balakot Formation to the north. The Patala Formation is characterized by a penetrative mylonitic foliation with associated down-dip fault striae and displays numerous shear-sense indicators, all suggestive of top-to-the-southwest thrust motion. Near the contact between the Patala and the Balakot formations, the Patala Formation occurs in thin imbricated slivers with the Balakot Formation, and displays sheared-out isoclinal folds. No primary sedimentary features are preserved in the Patala Formation. The contact region is therefore intensely tectonized and we believe that there is inadequate evidence for a conformable contact.

Thus, the petrographic and structural constraints are consistent with the detrital mica ages, showing that the Balakot Formation is no older than 36–40 Myr and is possibly significantly younger. Therefore, the first record of continental foreland basin sediments that contain evidence for erosion from the Himalayan metamorphic belt is $>15\text{--}20$ Myr later than previously believed; models influenced by the older age will need to be re-evaluated in the light of these new data.

Stratigraphic data provide one of the most compelling lines of evidence for the timing and diachrony of collision^{7,20}; initiation of orogen-derived foreland basin sedimentation provides a minimum age for orogenic loading of the crust. The marine to continental facies transition in the suture and foreland basin, and variation along strike, has been used to determine the timing and degree of diachrony of collision^{6–8}, from which estimates of the amount of accommodation of strain by extrusion have been made⁶.

Rowley⁶ concluded that only in western Zaskar–Hazara was dating (on the oldest syncollisional continental red beds of the Chulung La Formation in Zaskar and the Balakot Formation in Hazara) sufficiently precise to permit the accurate determination of the marine to continental facies transition at 51 Myr. Further east,

stratigraphic data can constrain collision only at <45 Myr, indicating >7 Myr diachrony. However, because the Chulung La Formation is unfossiliferous, its Early Eocene age is based only on its stratigraphic position, unconformably above the Late Paleocene Dibling Limestone, and its assumed but unproved correlation with the fossiliferous marine Kong Slates of that age^{21,22}. Hence, Early Eocene is only a robust maximum age for the Chulung La red beds, and the stratigraphic argument rests heavily on the Balakot Formation. Our revised Balakot Formation age of <35 Myr therefore removes all stratigraphic evidence for diachronous collision and also reduces the accuracy of the stratigraphically determined timing of collision (previously taken as 51 Myr)⁶ to between 55 Myr (ref. 21) (the youngest marine facies) and <35 Myr.

The earliest record of substantial Himalayan erosion, >15 Myr later than previously believed, forces a reconsideration of the timing and mechanisms of exhumation and uplift, and sedimentary responses to tectonic events. Although it is possible that older sediments are preserved further north beneath the thrust belt or near the base of the Indus or Bengal Fans, the Balakot Formation can no longer be taken as evidence of significant early erosional exhumation synchronous with the initial stages of collision, metamorphism and initiation of a period of rapid cooling, thrusting and extension that began in the Eocene^{9,10}. Thus, the relative importance of tectonic exhumation might be increased. Evidence of uplift by 55–50 Myr, as inferred from the interpreted thrust-belt barrier required to explain the contrasting provenance between the previously assumed coeval Chulung La suture zone sediments and the Balakot Formation foreland basin sediments¹¹, is no longer substantiated.

Our revision of the age and structure of the Balakot Formation affects current models of the early stages of foreland basin evolution. Burbank *et al.*¹² surmised that basin geometry was affected by changes in the rigidity of the underthrust plate with time. The thickness and age of the Balakot Formation supported the conten-

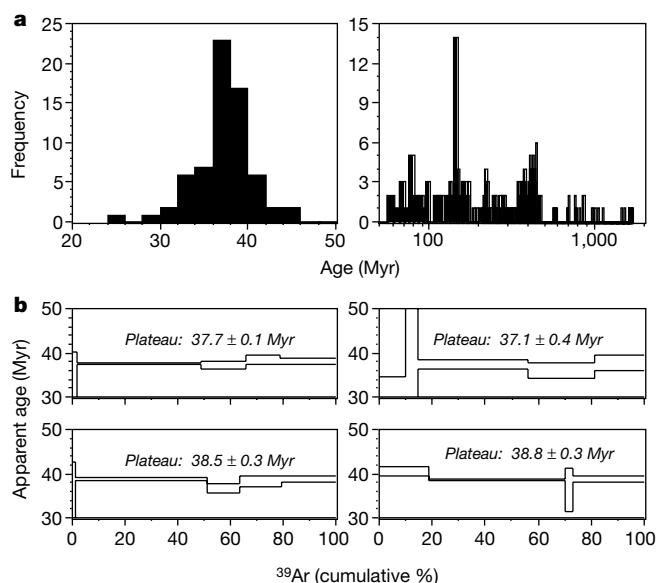


Figure 3 Summary of Ar–Ar analyses of single-crystal detrital white micas from the Balakot Formation sediments, Pakistan. **a**, Distribution of Ar–Ar laser-fusion and step-heating analyses of 257 individual white mica grains from the Balakot Formation, Pakistan. Grains 20–50 Myr old (left panel; 67 crystals) were sorted in 2-Myr-wide intervals; grains 50–2,000 Myr old (right panel; 190 crystals) were sorted in 5-Myr intervals; the analytical error is typically less than the bin width. **b**, Single-crystal Ar–Ar age spectra of four micas from a 36–40-Myr population; each step is drawn 2 s.d. tall. Analytical methods followed Richards *et al.*²⁴.

tion that younger and weaker Indian crust subducted during the earliest stages of collision would be reflected in a narrower and deeper basin. This early basin subsided rapidly, with sedimentation rates (calculated from the intercalated marl bands)⁵ up to an order of magnitude higher than during later stages of basin evolution in the Neogene¹². Our documentation of the structure of the Balakot Formation not only substantially decreases its stratigraphic thickness, and hence estimates of the depth of the basin at this time, but also shows that the marl bands cannot be used to determine sedimentation rates of the clastic deposits and therefore that there is no constraint on the early sedimentation and subsidence history of the basin. Our new age for the Balakot Formation negates its use as evidence of early orogenic loading and flexural subsidence, and removes the Hazara–Kashmir syntaxis region from its previously anomalous status to the otherwise basin-wide^{13,23} occurrence of a >20 Myr unconformity separating marine and continental facies. □

Received 6 April; accepted 11 December 2000.

- Ruddiman, W. F. & Kutzbach, J. E. Plateau uplift and climate change. *Sci. Am.* **264**, 42–50 (1991).
- Krishnaswami, S., Trivedi, J. R., Sarin, M., Ramesh, R. & Sharma, K. Strontium isotopes and rubidium in the Ganga–Brahmaputra rivers system: Weathering in the Himalaya, fluxes to the Bay of Bengal and contributions to the evolution of oceanic ⁸⁷Sr/⁸⁶Sr. *Earth Planet. Sci. Lett.* **109**, 2443–253 (1992).
- Richter, F. M., Rowley, D. B. & DePaolo, D. J. Sr isotopic evolution of seawater: the role of tectonics. *Earth Planet. Sci. Lett.* **109**, 11–23 (1992).
- Houseman, G. & England, P. Crustal thickening versus lateral expulsion in the Indian–Asian continental collision. *J. Geophys. Res.* **98**, 12233–12249 (1993).
- Bossart, P. & Ottiger, R. Rocks of the Murree Formation in Northern Pakistan: indicators of a descending foreland basin of late Palaeocene to middle Eocene age. *Eclogae Geol. Helv.* **82**, 133–165 (1989).
- Rowley, D. B. Age of initiation of collision between India and Eurasia: A review of stratigraphic data. *Earth Planet. Sci. Lett.* **145**, 1–13 (1996).
- Rowley, D. B. Minimum age of initiation of collision between India and Eurasia north of Everest based on the subsidence history of the Zhepure mountain section. *J. Geol.* **106**, 229–235 (1998).
- Uddin, A. & Lundberg, N. Cenozoic history of the Himalayan–Bengal system: Sand composition in the Bengal Basin, Bangladesh. *Geol. Soc. Am. Bull.* **110**, 497–511 (1998).
- Treloar, P. J., Rex, D. C. & Williams, M. P. The role of erosion and extension in unroofing the Indian Plate thrust stack, Pakistan Himalaya. *Geol. Mag.* **128**, 465–478 (1991).
- Treloar, P. J. Exhumation of high-grade Indian Plate rocks in North Pakistan: Mechanical implications of a multi-phase process. *Terra Nostra* **2**, 157–158 (1999).
- Critelli, S. & Garzanti, E. Provenance of the Lower Tertiary Murree redbeds (Hazara–Kashmir Syntaxis, Pakistan) and initial rising of the Himalayas. *Sedim. Geol.* **89**, 265–284 (1994).
- Burbank, D. W., Beck, R. A. & Mulder, T. The Himalayan foreland basin. In *The Tectonic Evolution of Asia* (eds Yin, A. & Harrison, T. M.) 149–188 (Cambridge Univ. Press, 1996).
- Najman, Y. M. R., Pringle, M. S., Johnson, M. R. W., Robertson, A. H. F. & Wijbrans, J. R. Laser ⁴⁰Ar/³⁹Ar dating of single detrital muscovite grains from early foreland basin sediments in India: Implications for early Himalayan evolution. *Geology* **25**, 535–538 (1997).
- Pivnik, D. A. & Wells, N. A. The transition from Tethys to the Himalaya as recorded in northwest Pakistan. *Geol. Soc. Am. Bull.* **108**, 1295–1313 (1996).
- Clauer, N. & Chaudri, S. Isotopic dating of very low-grade metasedimentary and metavolcanic rocks: techniques and methods. In *Low Grade Metamorphism* (eds Frey, M. & Robinson, D.) 202–226 (Blackwell Science, 1999).
- Bossart, P. *Eine Neuinterpretation der Tektonik der Hazara Kashmir Syntaxis (Pakistan)*. Thesis, ETH Zurich, Switzerland (1986).
- Kubler, B. La cristallinité de l'illite à les zones tout a fait supérieures du métamorphisme. In *Étages Tectoniques* 105–122. (A la Baconnière, Neuchâtel, Switzerland, 1967).
- Blenkinsop, T. G. Definition of low-grade metamorphic zones using illite crystallinity. *J. Met. Petrol.* **6**, 623–636 (1988).
- Weber, K. Notes on the determination of illite crystallinity. *Neues Jb. Miner. Monatshefte* **6**, 267–276 (1972).
- Butler, R. When did India hit Asia? *Nature* **373**, 20–21 (1995).
- Garzanti, E., Baud, A. & Mascle, G. Sedimentary record of the northward flight of India and its collision with Eurasia (Ladakh Himalaya, India). *Geodinim. Acta* **1**, 297–312 (1987).
- Garzanti, E., Critelli, S. & Ingersoll, R. V. Paleogeographic and paleotectonic evolution of the Himalayan range as reflected by detrital modes of Tertiary sandstones and modern sands (Indus transects, India and Pakistan). *Geol. Soc. Am. Bull.* **108**, 631–642 (1996).
- DeCelles, P. G., Gehrels, G. E., Quade, J. & Ojha, T. P. Eocene–early Miocene foreland basin development and the history of Himalayan thrusting, western and central Nepal. *Tectonics* **17**, 741–765 (1998).
- Richards, J. P., Noble, S. R. & Pringle, M. S. A revised Late Eocene age for porphyry Cu magmatism in the Escondida Area, Northern Chile. *Econ. Geol.* **94**, 1231–1248 (1999).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank E. Laws for field assistance, I. Hussain and Banares for driving, and Major Saeed for his co-operation during major road building; R. Marr, J. Nicholls and M. Stout for

assistance with the electron probe analyses; B. Davidson and J. Imlach for assistance with the Ar–Ar analyses; A. Calder for performing the illite crystallinity analyses; N. Portelance for drafting some of the figures; and D. Burbank and P. Copeland for critical reviews. This work was funded by a Royal Society Dorothy Hodgkin Fellowship and completed during a Royal Society International Fellowship to Y.N. Additional support was provided by the Fold–Fault Research Project at the University of Calgary and a Royal Society Research Grant to Y.N. Ar–Ar analyses were funded by NERC support of the Argon Isotope Facility at the Scottish Universities Environmental Research Centre (SUERC).

Correspondence and requests for materials should be addressed to Y.N. (e-mail: y.najman@glg.ed.ac.uk).

Evidence for mantle metasomatism by hydrous silicic melts derived from subducted oceanic crust

Gaelle Prouteau^{*†}, Bruno Scailliet[‡], Michel Pichavant[‡] & René Maury^{*}

^{*} UMR 6538, UBO, 6 avenue Le Gorgeu, P.B. 809, 29285 Brest, France

[‡] UMR G113, ISTO, 1a rue de la Férollerie, 45071 Orléans, France

The low concentrations of niobium, tantalum and titanium observed in island-arc basalts are thought to result from modification of the sub-arc mantle by a metasomatic agent, deficient in these elements, that originates from within the subducted oceanic crust¹. Whether this agent is an hydrous fluid² or a silica-rich melt³ has been discussed using mainly a trace-element approach⁴ and related to variable thermal regimes of subduction zones⁵. Melting of basalt in the absence of fluid both requires high temperatures and yields melt compositions unlike those found in most modern or Mesozoic island arcs^{6,7}. Thus, metasomatism by fluids has been thought to be the most common situation. Here, however, we show that the melting of basalt under both H₂O-added and low-temperature conditions can yield extremely alkali-rich silicic liquids, the alkali content of which increases with pressure. These liquids are deficient in titanium and in the elements niobium and tantalum and are virtually identical to glasses preserved in mantle xenoliths found in subduction zones⁶ and to veins found in exhumed metamorphic terranes of fossil convergent zones⁷. We also found that the interaction between such liquids and mantle olivine produces modal mineralogies that are identical to those observed in metasomatized Alpine-type peridotites⁸. We therefore suggest that mantle metasomatism by slab-derived melt is a more common process than previously thought.

A mid-ocean ridge basalt from the Juan de Fuca ridge (Table 1) was reacted between 10 and 30 kbar with bulk H₂O contents of 2 to 10 wt% (see ref. 9 for methods), that is, largely in excess of the H₂O content of metamorphosed basalts, and at temperatures below 1,000 °C. Recent experimental work has shown that melting under these conditions (addition of H₂O, low temperatures) is required to account for the geochemical and physical characteristics of slab melts⁹, and approaches the most likely pressure–temperature (*P–T*) conditions for slab melting (750 °C, 30 kbar) as constrained from thermal modelling¹⁰.

Representative glass (that is, quench melt) compositions are listed in Table 1. In an An–Ab–Or projection (Fig. 1), the glass compositions define a trend parallel to the albite–anorthite join. Glasses

[†] Present address: Laboratoire de Pétrologie, UPMC, 4 Place Jussieu, 75252, Paris, France.

produced at both 10 and 20 kbar cluster in the tonalitic field, whereas those obtained at 30 kbar fall within the trondhjemite field. The latter have Na₂O contents in excess of 7 wt%, increasing as temperature decreases, and are clearly peralkaline (molar Al₂O₃/Na₂O+K₂O+CaO<1), whereas all the glasses produced at lower pressures tend to be either metaluminous or slightly peraluminous. The reason for the large compositional jump observed between 10–20 and 30 kbar is that amphibole breaks down in this pressure range, and the modal abundances of garnet and clinopyroxene increase. The Na₂O released by the amphibole breakdown is not fully balanced by the increase in the Na content of clinopyroxene (that is, the jadeite component) and thus, for a given *T* and bulk H₂O content, the liquid becomes more sodic and less calcic as pressure increases beyond the amphibole stability field. Extensive garnet crystallization at 30 kbar also lowers the Al₂O₃ content of the glasses relative to those at 20 kbar (Table 1). In the CaO versus Na₂O diagram (Fig. 2), all glasses so far produced in H₂O-added melting experiments on basalts up to 30 kbar^{11–13} display a broad negative trend that completely overlaps that field of natural glasses or veins believed to be slab melts^{6,7}, in contrast to results from dehydration melting experiments^{13–15} (Fig. 2b). A similar feature is observed for other elements such as FeO or TiO₂. The melt TiO₂ content (Table 1) agrees with existing rutile solubility models¹⁶. At pressures lower

than 10 kbar the glasses are characterized by high CaO and relatively low Na₂O contents, as well as a high Al₂O₃ content, due to the lack of both plagioclase and garnet in the residue^{12,13}. In contrast, glasses obtained in the pressure range 20–30 kbar coexist with an eclogitic residue. Results obtained at 30 kbar match the composition of metasomatic veins the most closely, in particular the extreme Na₂O enrichment. This suggests that these experimental conditions—temperatures below 900 °C and bulk water content around 5–6 wt%, thus requiring the presence of a fluid phase before melting—reproduce those of vein production. The origin of free water may lie in the dehydration of serpentinite beneath the oceanic crust of the subducted plate^{17,18}. The water thus produced is expected to dissolve in the melt, except at low melt fraction where there may be enough water to saturate the melt. Clearly, these conditions of slab melting better approach those predicted from thermal modelling¹⁰ than do the conditions of the dehydration melting experiments. Therefore, we conclude that free water is needed at depth for slab melting to occur in present-day subduction zones. Experimental constraints on slab melts predict meltwater contents of at least 15 wt% (ref. 9), whereas mass balance calculations on the run products obtained at 30 kbar yield meltwater contents between 24 and 30 wt%. The unusual major-element composition of slab melts appears to be more characteristic of their origin than their trace element signature. In particular, melting in the garnet field of basalt underplated beneath the continental crust may yield silicic melts with trace element patterns similar to those of slab melts¹⁹, but they will be granitic instead of trondhjemitic because less water is available in the lower crust and dehydration melting reactions are favoured.

Interaction of such trondhjemitic melts with the overlying mantle wedge has been experimentally simulated at 15 kbar, as in previous similar works^{20–22}, by mixing a synthetic equivalent of the trondhjemite melt obtained during melting of the basalt with a forsteritic olivine (Fo 90), the dominant mineral in peridotites (Table 2). Liquid and mineral compositions are listed in Table 2. At 1,000 °C, an orthopyroxene+liquid assemblage is produced. At 900 °C, in addition to orthopyroxene, both an amphibole and phlogopite are reaction products of the interaction process. At both temperatures, olivine appears not to be stable, although residual crystals persist. Glass compositions are notably more magnesian and silica-poor than the starting trondhjemite, being dacitic instead of rhyolitic, yet their strongly alkali-rich character still persists or even increases (Table 2). An increase of the Mg/(Mg+Fe) ratio of the liquid together with massive orthopyroxene crystallization are key features of geochemical models of slab melt/peridotite interaction²³ and reproduce diagnostic petrographic characteristics of slab-derived magmas²⁴. In addition, the compositions of amphibole and phlogopite, the two main minerals diagnostic of mantle

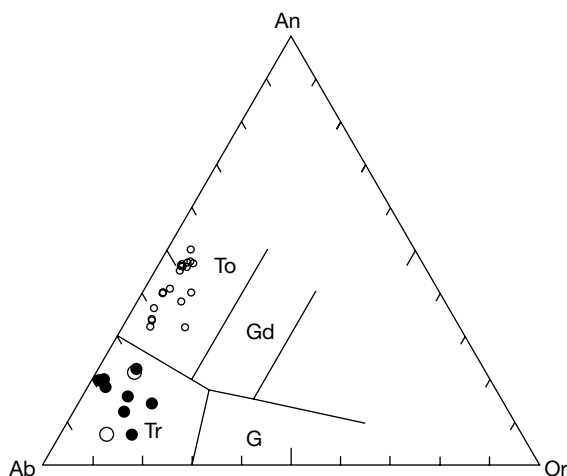


Figure 1 Normative (CIPW) compositions of natural slab melts (black circles) and glasses obtained in melting experiments of basalts with added H₂O (this study). Small white circles represent experiments done at or below 20 kbar and large white circles experiments at 30 kbar. An, anorthite; Ab, albite; Or, orthoclase; To, tonalite, Tr, trondhjemite; Gd, granodiorite; G, granite.

Table 1 Experimental and natural glass/vein compositions and experimental phase proportions

	P (kbar)	T (°C)	H ₂ O*	SiO ₂	TiO ₂	Al ₂ O ₃	FeO†	MgO	CaO	Na ₂ O	K ₂ O	Gl	Hbl	Cpx	Gt	Tmt	Ilm	Rut	Sph
Experimental glasses from hydrous basalt melting																			
Starting‡	–	–	–	49.9	2.3	13.7	12.7	6.8	10.6	2.5	0.3	–	–	–	–	–	–	–	–
27	10	841	9.9	69.8	0.2	18.7	0.9	0.5	5.8	3.4	0.6	41.1	26.0	21.7	–	8.6	2.6	–	–
38	20	800	9.2	68.8	0.3	16.7	2.9	1.0	5.9	3.4	0.7	28.0	67.4	–	–	–	0.1	–	7.0
39	20	900	10.2	65.4	0.4	20.9	1.7	0.2	7.0	3.7	0.6	36.2	39.1	20.9	–	–	4.0	–	–
40	20	1,000	10.0	59.3	1.5	20.1	5.1	0.9	7.5	4.3	0.8	46.8	26.5	20.8	–	5.9	–	–	–
41	30	900	6.4	70.9	0.3	16.1	0.5	0.1	1.4	9.3	1.5	23.9	–	42.9	31.8	–	0.4	0.9	–
42	30	1,000	6.0	68.0	0.6	17.9	1.2	0.2	3.7	7.1	1.1	25.6	–	34.8	38.4	–	–	1.1	–
Felsic veins in mantle xenoliths or in metamorphic complexes from subduction zones																			
Val55/4	–	–	–	67.8	0.4	18.0	0.6	0.2	1.1	7.9	2.1	–	–	–	–	–	–	–	–
Val55/4	–	–	–	63.9	0.9	19.7	2.7	1.0	3.0	5.7	0.9	–	–	–	–	–	–	–	–
41867A	–	–	–	72.2	0.1	16.5	0.2	0.1	2.8	6.6	0.3	–	–	–	–	–	–	–	–

Analyses and phase proportions given in wt%. Glass analyses normalized to 100% anhydrous. Gl, glass; Hbl, hornblende; Cpx, clinopyroxene; Gt, garnet; Tmt, titanomagnetite; Ilm, ilmenite; Rut, rutile; Sph, sphene. Run durations are between 25 h at 1,000 °C to 144 h at 841 °C, see ref. 9 for methods. Glass compositions Val55/4 are from ref. 6. The pegmatite vein 41867A is from ref. 7.

* H₂O loaded to the capsule.

† Total Fe calculated as FeO.

‡ Dry starting glass composition used in crystallization experiments.

metasomatism, are remarkably similar to those of natural amphibole and phlogopite found in some Alpine-type peridotite complexes believed to have been pervasively metasomatized by infiltrating slab-derived melts⁸. In particular, both minerals are characterized by high Mg/(Fe+Mg) ratios and display strong enrichment in Na₂O. Given the large number of parameters involved in the interaction process (that is, compositions of peridotite and infiltrating melt, solid/melt ratio, pressure and temperature of the interaction) it is expected that metasomatic phases display a wide range of composition, as in previous laboratory studies^{20–22} and natural occurrences²⁵. However, the sodic character of metasomatic minerals can be related to the sodic character of the metasomatizing agent, as evidenced by the trondhjemite veins of Kamchatka. Variable Na₂O enrichment of metasomatic minerals would trace the more or less trondhjemitic character of the percolating melt which in turn depends on the conditions of its genesis, in particular pressure, as shown above. The similarity between the compositions of experimental and natural phases suggests that the interaction experiments reproduce the conditions of mantle metasomatism and in particular the chemistry of the metasomatic agent.

The above experimental evidence bears direct constraints on metasomatic processes in arc settings. The low-temperature, fluid-present melting conditions above suggest that slab melting can occur in presently subducting plates, as previously suggested on geochemical grounds⁵. Although slab melting was originally supposed to have occurred mostly during the Archaean²⁶ or only when young and hot plate subducts in modern settings⁵, the increasing number of occurrences of magmas with slab melt signatures worldwide²⁷ suggests that conditions for slab melting are realized more often than currently believed. The slab melt contribution as a metasomatizing agent of the sub-arc mantle may thus be important. Mass balance calculations indicate that 15 g of trondhjemite yields 1 g of phlogopite upon reaction with peridotite, whereas 180 g of hydrous fluid are needed to precipitate an equivalent mass of phlogopite during a peridotite–fluid interaction²⁸. Thus, as far as metasomatic mineral production is concerned, slab-derived melts are about ten times more efficient than hydrous fluids. Consequently, the geochemical imprint of slab melts can be important in modern arc settings even if slab melting is probably less common than slab dehydration²⁹. In particular, calculated trace-element concentrations in slab melts in equilibrium with an eclogitic residue (garnet+clinopyroxene+/-rutile) are enriched in light-rare-earth and large-ion-lithophile elements and depleted in high-field-strength elements (La/Nb = 7; Ba/Nb = 186; ref. 22) and will

impart their geochemical signature to magmas generated in the mantle wedge²⁴.

The overall slab contribution to the source of arc magmas is estimated not to exceed 3% on the basis of oxygen isotope systematics of oceanic arc lavas³⁰. The reason why slab melts (adakites) are found in volumetrically minor amounts relative to regular calc-alkaline arc magmas may be either physical or chemical. Silica-rich liquids can equilibrate with mantle minerals provided

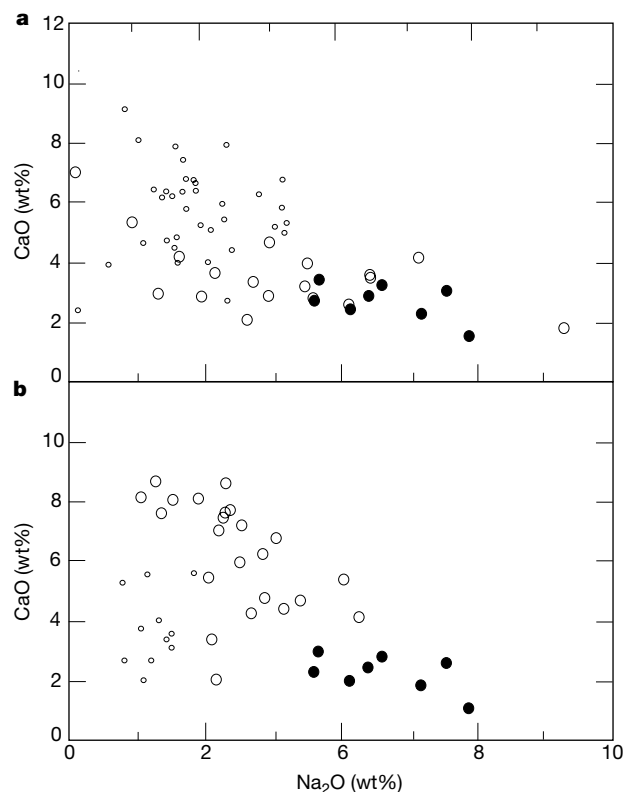


Figure 2 Na₂O versus CaO plot of natural slab melts^{6,7} (black circles) and of glasses obtained in melting experiments (large white circles) using N-MORB type basalt protoliths. **a**, Glasses obtained in melting experiments with added H₂O (this study and refs 11–13). **b**, Glasses obtained in dehydration melting experiments^{13–15}. Small white circles represent experiments done at or below 10 kbar (in both **a** and **b**).

Table 2 Glass and mineral compositions from trondhjemite–Fo90 olivine interactions at 15 kbar

	SiO ₂	TiO ₂	Al ₂ O ₃	FeO*	MgO	CaO	Na ₂ O	K ₂ O	NiO	total
Starting end members										
Trondh	71.2	0.5	16.9	0.5	0.1	1.5	7.8	1.5	–	100
Olivine	40.8	–	0.1	10.8	47.8	–	–	–	0.3	99.7
Run product compositions										
1,000 °C, 7.8 wt% H ₂ O†										
Gl	65.9	0.5	17.6	1.0	2.9	1.4	8.9	1.6	<0.1	92.2
Opx	58.3	0.1	0.8	3.8	35.3	0.9	0.1	–	0.2	99.5
900 °C, 9.6 wt% H ₂ O										
Gl	64.2	0.5	18.8	1.1	1.5	1.0	11.4	1.5	<0.1	89.6
Opx	57.7	0.1	0.8	6.0	33.4	0.9	0.2	–	0.2	99.4
Phl	41.9	1.6	14.2	2.9	24.4	–	2.7	6.3	0.2	94.4
Amph	50.9	1.7	6.1	3.5	20.6	8.9	4.2	0.3	0.1	96.5
Hydrous phases in metasomatised peridotite‡										
Phl	41.9	0.7	14.3	2.8	25.3	0.0	1.6	6.5	–	94.6
Amph	49.9	0.4	7.5	3.4	20.2	10.2	3.8	0.6	–	98.3

Mixtures consisted in 60 wt% trondhjemite dry glass and 40 wt% olivine. Analyses given in wt%. Glass analyses normalized to 100% anhydrous. Gl, glass; Opx, orthopyroxene; Phl, phlogopite; Amph, amphibole; trondh, trondhjemite.

* Total Fe calculated as FeO.

† Bulk H₂O content of the charge and temperature of the interaction experiment.

‡ Phlogopite and amphibole from the Finero peridotite massif⁸.

they are alkali-rich³¹. This alkali effect is most pronounced at pressures below 20 kbar, however. At 30 kbar, liquids in equilibrium with mantle peridotite are not alkali-rich. The alkali content of experimental slab melts increases with pressure, and comparison with their natural equivalents has shown that slab melting is likely to occur at 30 kbar or more. Thus, besides the physical aspects of magma transport and percolation, any alkali-rich slab melt produced at or beyond 30 kbar, as in our experiments, will be in strong chemical disequilibrium with the overlying mantle wedge. Extensive interaction between these melts and peridotite is expected to occur and the alkali-rich, silica-rich melts will have little opportunity to reach shallower levels unchanged. On this basis, preservation or eruption of genuine slab melts can be anticipated to be difficult, accounting for their relative rarity in most arcs. The chemical differences observed between experimentally produced slab-melts and those erupted or preserved as melt inclusions³² must in fact reflect various extents of interaction of the latter with mantle material during upward migration. □

Received 22 June 2000; accepted 16 January 2001.

- Hofmann, A. W. Chemical differentiation of the Earth: the relationships between mantle, continental crust, and oceanic crust. *Earth Planet. Sci. Lett.* **90**, 297–314 (1988).
- Eiler, J. M., McInnes, B., Valley, J. W., Graham, C. M. & Stolper, E. M. Oxygen isotope evidence for slab-derived fluids in the sub-arc mantle. *Nature* **393**, 777–781 (1998).
- Nicholls, I. A. & Ringwood, A. E. Effect of water on olivine stability in tholeiites and production of silica saturated magmas in island arc environment. *J. Geol.* **81**, 285–300 (1973).
- Hawkesworth, C. J., Turner, S. P., McDermott, E., Peate, D. W. & van Calsteren, P. U–Th isotopes in arc magmas: implications for element transfer from the subducted crust. *Science* **276**, 551–555 (1997).
- Defant, M. J. & Drummond, M. S. Derivation of some modern arc magmas by melting of young subducted lithosphere. *Nature* **347**, 662–665 (1990).
- Kepezhinskis, P. K., Defant, M. J. & Drummond, M. S. Na metasomatism in the island-arc mantle by slab melt-peridotite interaction: evidence from mantle xenoliths in the north Kamchatka arc. *J. Petrol.* **36**, 1505–1527 (1995).
- Sorensen, S. S. & Grossman, J. N. Enrichment of trace elements in garnet amphibolites from a paleo-subduction zone: Catalina Schist, southern California. *Geochim. Cosmochim. Acta* **53**, 3155–3177 (1989).
- Zanetti, A., Mazzucchelli, Rivalenti, G. & Vannucci, R. The Finero phlogopite-peridotite massif: an example of subduction-related metasomatism. *Contrib. Mineral. Petrol.* **134**, 107–122 (1999).
- Prouteau, G., Scaillet, B., Pichavant, M. & Maury, R. C. Fluid-present melting of ocean crust in subduction zones. *Geology* **27**, 1111–1114 (1999).
- Peacock, S. M., Rushmer, T. & Thompson, A. B. Partial melting of subducting oceanic crust. *Earth Planet. Sci. Lett.* **121**, 227–244 (1994).
- Whitner, K. T. & Newton, R. C. Experimental melting of hydrous low-K tholeiite: evidence on the origin of Archean cratons. *Bull. Geol. Soc. Denmark* **39**, 213–228 (1991).
- Spulber, S. D. & Rutherford, M. J. The origin of rhyolite and plagiogranite in oceanic crust: an experimental study. *J. Petrol.* **24**, 1–25 (1983).
- Beard, J. S. & Lofgren, G. E. Dehydration melting and water-saturated melting of basaltic and andesitic greenstones and amphibolites at 1, 3, and 6.9 kb. *J. Petrol.* **32**, 365–401 (1991).
- Rapp, R. P. & Watson, E. B. Dehydration melting of metabasalt at 8–32 kbar: implications for continental growth and crust-mantle recycling. *J. Petrol.* **36**, 891–931 (1995).
- Sen, C. & Dunn, T. Dehydration melting of a basaltic composition amphibolite at 1.5 and 2.0 GPa: implications for the origin of adakites. *Contrib. Mineral. Petrol.* **117**, 394–409 (1994).
- Ryerson, F. J. & Watson, E. B. Rutile saturation in magmas: implications for Ti-Nb-Ta depletion in island-arc basalts. *Earth Planet. Sci. Lett.* **86**, 225–239 (1987).
- Ulmer, P. & Trommsdorff, V. Serpentine stability to mantle depths and subduction-related magmatism. *Science* **268**, 858–861 (1995).
- Schmidt, M. & Poli, S. Experimentally based water budget for dehydrating slabs and consequences for arc magma generation. *Earth Planet. Sci. Lett.* **163**, 361–379 (1998).
- Atherton, M. P. & Petford, N. Generation of sodium-rich magmas from newly underplated basaltic crust. *Nature* **362**, 144–146 (1993).
- Carroll, M. J. & Wyllie, P. J. Experimental phase relations in the system peridotite-tonalite-H₂O at 15 kbar: implications for assimilation and differentiation processes at the crust-mantle boundary. *J. Petrol.* **30**, 1351–1382 (1989).
- Sen, C. & Dunn, T. Experimental modal metasomatism of a spinel lherzolite and the production of amphibole-bearing peridotite. *Contrib. Mineral. Petrol.* **119**, 422–432 (1994).
- Rapp, R. P., Shimizu, N., Norman, M. D. & Applegate, G. S. Reaction between slab-derived melts and peridotite in the mantle wedge: experimental constraints at 3.8 GPa. *Chem. Geol.* **160**, 335–356 (1999).
- Kelemen, P., Shimizu, N. & Dunn, T. Relative depletion of niobium in some arc magmas and the continental crust: partitioning of K, Nb, La, and Ce during melt/rock reaction in the upper mantle. *Earth Planet. Sci. Lett.* **120**, 111–134 (1993).
- Yogodzinski, G. M., Volynets, O. N., Koloskov, A. V. & Seliverstov, N. I. Magnesian andesites and the subduction component in a strongly calcalkaline series at Piip volcano, far western Aleutian. *J. Petrol.* **34**, 163–204 (1994).
- Ionov, D. A. & Hofmann, A. W. Nb-Ta-rich mantle amphiboles and micas: implications for subduction-related metasomatic trace element fractionations. *Chem. Geol.* **131**, 341–356 (1995).
- Martin, H. Effect of steeper Archean geothermal gradient on geochemistry of subduction zone magmas. *Geology* **14**, 753–756 (1986).
- Drummond, M. S., Defant, M. J. & Kepezhinskis, P. K. Petrogenesis of slab derived-tonalite-dacite adakite magmas. *Trans. R. Soc. Edinburgh* **87**, 205–215 (1996).

- Schneider, M. E. & Eggler, D. H. Fluids in equilibrium with peridotite minerals: implications for mantle metasomatism. *Geochim. Cosmochim. Acta* **50**, 711–724 (1986).
- Tatsumi, Y. Migration of fluid phases and genesis of basalt magmas in subduction zones. *J. Geophys. Res.* **94**, 4697–4707 (1989).
- Eiler, J. M. *et al.* Oxygen isotopes geochemistry of oceanic-arc lavas. *J. Petrol.* **41**, 229–256 (2000).
- Hirschman, M. M., Baker, M. B. & Stolper, E. M. The effect of alkalis on the silica content of mantle-derived melts. *Geochim. Cosmochim. Acta* **62**, 883–902 (1998).
- Schiano, P. *et al.* Hydrous, silica-rich melts in the sub-arc mantle and their relationship with erupted arc-lavas. *Nature* **377**, 595–600 (1995).

Correspondence and requests for materials should be addressed to B.S. (e-mail: bscaillie@cnrs-orleans.fr).

Branched integumental structures in *Sinornithosaurus* and the origin of feathers

Xing Xu*, Zhong-he Zhou* & Richard O. Prum†

* Institute of Vertebrate Paleontology and Paleoanthropology, Academia Sinica, PO Box 643, Beijing, 100044, People's Republic of China

† Department of Ecology and Evolutionary Biology, and Natural History Museum, University of Kansas, Lawrence, Kansas 66045, USA

The evolutionary origin of feathers has long been obscured because no morphological antecedents were known to the earliest, structurally modern feathers of *Archaeopteryx*¹. It has been proposed that the filamentous integumental appendages on several theropod dinosaurs are primitive feathers^{2–6}; but the homology between these filamentous structures and feathers has been disputed^{7–9}, and two taxa with true feathers (*Caudipteryx* and *Protarchaeopteryx*) have been proposed to be flightless birds^{8,10}. Confirmation of the theropod origin of feathers requires documentation of unambiguously feather-like structures in a clearly non-avian theropod. Here we describe our observations of the filamentous integumental appendages of the basal dromaeosaurid dinosaur *Sinornithosaurus millenii*, which indicate that they are compound structures composed of multiple filaments. Furthermore, these appendages exhibit two types of branching structure that are unique to avian feathers: filaments joined in a basal tuft, and filaments joined at their bases in series along a central filament. Combined with the independent phylogenetic evidence supporting the theropod ancestry of birds^{11–13}, these observations strongly corroborate the hypothesis that the integumental appendages of *Sinornithosaurus* are homologous with avian feathers. The plesiomorphic feathers of *Sinornithosaurus* also conform to the predictions of an independent, developmental model of the evolutionary origin of feathers¹⁴.

Sinornithosaurus millenii (Fig. 1) is a non-avian, basal dromaeosaurid dinosaur from the Lower Cretaceous Yixian formation (~124.6 Myr ago), Liaoning, China⁵. Along with *Sinosauropteryx*³ (a basal coelurosaur), *Beipiaosaurus*⁴ (a therizinosauroid) and *Microraptor*⁶ (a dromaeosaur), *Sinornithosaurus*⁵ has many filamentous integumental structures that may be plesiomorphic feathers. As in *Beipiaosaurus* and *Microraptor*, the filamentous integumental appendages of *Sinornithosaurus* are indistinguishable from the contour feathers of birds preserved in the same deposits (compared with holotypes of *Confuciusornis sanctus*, *Eoenantiornis buhler* and *Chanchengornis hengdaoziensis*). In these avian specimens, the indisputable contour feathers on the body are preserved as diffuse arrays of filaments of varying diameters without a prominent rachis or central shaft. Despite these similarities, homology between the

integumental filaments of *Sinornithosaurus* and avian feathers has been questioned⁹.

Further preparation of the holotype of *Sinornithosaurus* (IVPP V12811) has revealed new details about the morphology of its filamentous integumental appendages, which are preserved as carbonized filaments around the specimen (Fig. 1). The appendages are generally 30–45 mm long and 1–3 mm wide. Many appendages remain closely associated with the integument of the skull, neck, forelimbs, legs and tail. Others were dissociated from the integument and are preserved in the matrix around the specimen. No remiges or rectrices (that is, flight feathers) with prominent rachis have been observed. These appendages are unlikely to be collagenous dermal or musculoskeletal structures⁷ because they are clearly preserved as external integumental structures. Furthermore, they are too numerous and broadly distributed to be partially ossified tendons.

Avian feathers are characterized by a complex, branched structure of keratinaceous filaments that grow by a unique mechanism from cylindrical feather follicles^{14,15}. This branching structure is the most distinctive morphological feature of feathers. Our observations indicate that there are similar branching structures in the integumental appendages of *Sinornithosaurus*. Generally, the *Sinornithosaurus* appendages are preserved not as solitary fibres but as compound structures containing multiple filaments (Fig. 2). This compound filamentous structure is particularly notable in appendages that were dissociated from the body during preservation and deposited around the specimen (Fig. 2). The preservation of numerous, randomly oriented, compound structures overlying each other shows that these integumental appendages were dissociated from the skin and from one another before preservation, but that the component filaments within each appendage remained associated (Fig. 2), indicating that the component filaments in each appendage were joined together. In contrast, single, unbranched, independent, hair-like integumental filaments would be unlikely to be preserved in similarly sized groups of overlying filaments. Multiple hair-like filaments from independent integumental structures might remain associated if they were attached to a single piece of skin that sloughed off the body during decomposition, but the varying orientation of the dissociated, overlying *Sinornithosaurus* appendages (Fig. 2) argues directly against this interpretation.

In addition to their compound structure, the integumental appendages of *Sinornithosaurus* exhibit two types of feather-like

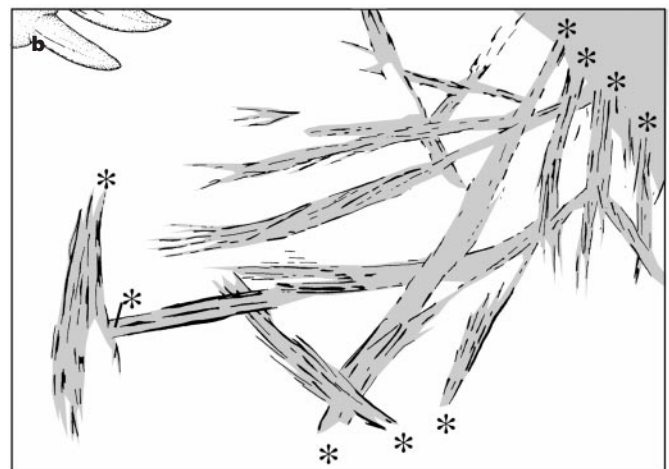


Figure 2 Filamentous integumental appendages of *S. millenii* dissociated from the integument and preserved overlying each other in the matrix near the skull. **a**, Picture of the preparation. Scale bar, 5 mm. **b**, Illustrated reconstruction of the appendages showing the positions of the observed filaments (lines) and the inferred outline of the appendages (shading). Asterisks, the proximal ends of each appendage. Each appendage is composed of multiple component filaments.

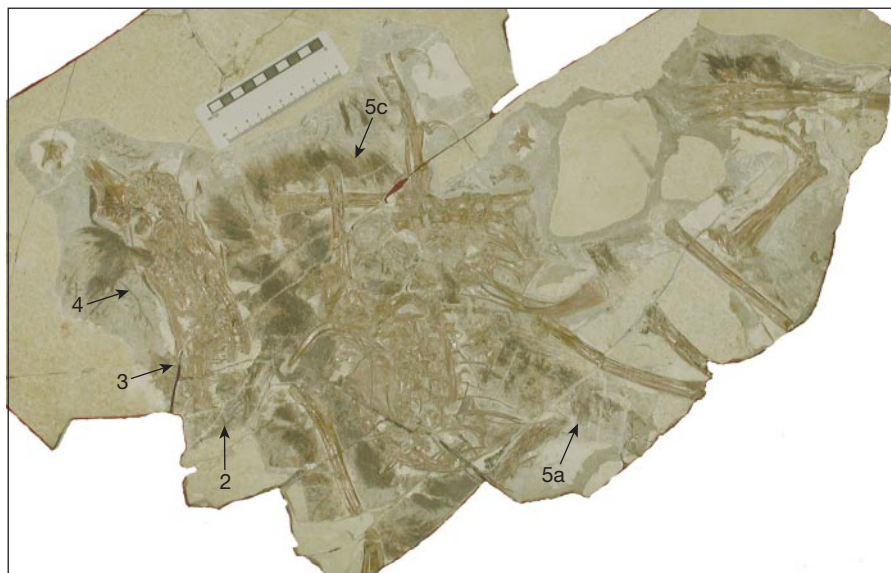


Figure 1 *Sinornithosaurus millenii*, holotype (IVPP 12811). Carbonized, filamentous, integumental appendages, 30–45 mm long, are attached to the skull, forelimbs, tail and other skeletal elements. Dissociated appendages are preserved in the substrate around

the specimen. The locations of the appendages illustrated in subsequent Figures are labelled with the corresponding Figure numbers.

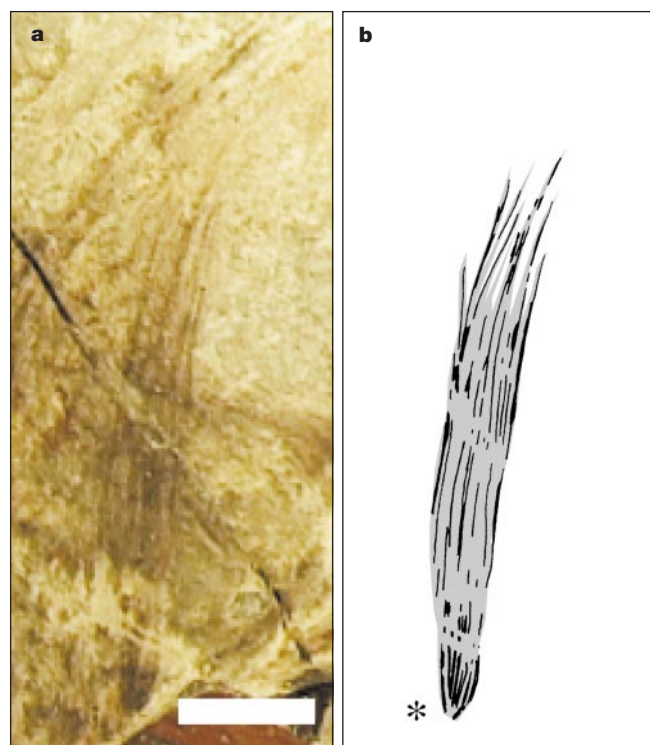


Figure 3 Filamentous integumental appendage of *S. millenii* from the dorsal surface of the snout. **a**, Picture of the preparation. Scale bar, 5 mm. **b**, Illustrated reconstruction of the appendage showing the positions of the observed filaments (lines) and the inferred outline of the appendage (shading). Asterisk, the proximal end of the appendage. The nearly parallel fibres converge at the base on a single point.

branching structure. In several instances, component filaments are generally similar in length, extensively parallel for most of their length, and converge conspicuously at their bases on a single point (Fig. 3). In this form of branched structure, many unbranched filaments converge on a single point in a prominent tuft.

One dissociated integumental appendage is preserved in a curved position that demonstrates a second form of branched structure (Fig. 4). Here, the distal tips of the component filaments occur along the entire length of the appendage, and the filaments form the margins of the appendage in a manner similar to a pennaceous feather vane. Furthermore, the component filaments are not parallel to the central axis of the appendage as it curves. Rather, filaments from both sides are straighter than the appendage as whole, and each filament converges at its base on the central midline of the composite appendage. This conformation would be expected only for filaments that were joined at their bases to other filaments or to a central filament (or shaft). These observations are not consistent with a frayed, fibrous composite material (for example, ossified tendon) because the appendage remains essentially the same width throughout its length even though new filaments are consistently added to it. The distal filaments could not continue all the way to the base of the appendage at their observed diameters or the width of the appendage would increase substantially as new filaments were added. As the component filaments are not as long as the appendage itself, it appears that they are joined at their bases to another filament within the appendage.

In several instances, there are additional indications of a larger central filament and secondary branches within a single appendage. Some appendages display a general outline of the margins of the appendage with a substantially larger, darker, more prominent filament running down the centre (Fig. 5a, b). In one instance, a single filament can be seen joining at its base to the larger central



Figure 4 Dissociated filamentous integumental appendage of *S. millenii* from near the skull. **a**, Arrows indicate the distal tips of some component filaments. Scale bar, 5 mm. **b**, Illustrated reconstruction of the appendage showing the positions of the observed filaments (lines) and the inferred outline of the appendage (shading). Asterisk, the proximal end of the appendage. The curved position of the appendage reveals its compound structure. Each filament converges on the centre of the appendage at its base.

filament at an acute angle (Fig. 5a, b), in the manner of the rachis and barbs of a pinnate feather. The filamentous integumental appendages along the right ulna are preserved as a substantial mat of filaments (Fig. 5c). The heterogeneous filaments differ strikingly in diameter and in angle to the ulna, consistent with impressions of many fibres from different but closely associated branched appendages. The longest, darkest and thickest filaments are straight, perpendicular to the skin, and arrayed along the ulna at regular distances (around 0.5 mm apart). These larger filaments are also largest at their bases.

The integumental appendages of *Sinornithosaurus* are compound structures composed of multiple filaments with two types of branched structure. First, the tufts of filaments joined at their bases are identical in structure to avian natal down feathers in which multiple filamentous barbs are basally fused to a single calamus¹⁵. Among modern vertebrate integumental appendages, this type of branched structure is unique to avian feathers¹⁵. (Chinchilla (Chinchillidae) hair is also tufted, but each filament actually grows from an independent root within a follicle^{16,17}.) Second, the serial branching of filaments along a central shaft is identical in structure to the barbs and the rachis of a pennaceous feather, and is also unique to avian feathers. The integumental appendages of *Sinornithosaurus* are different from most modern avian feathers in their apparent lack of barbules. Thus, *Sinornithosaurus* appendages could not have formed a closed pennaceous vane.

The compound filamentous structure and the two types of feather-like branching in the integumental appendages of *Sinornithosaurus* strongly indicate that these structures are homologous with avian feathers. A previous phylogenetic analysis⁵ indicates that *Sinornithosaurus* is not a bird but is a basal lineage of the dromaeosaurids which, by themselves or in combination with troodontids, have been repeatedly considered to be the lineage of theropods most closely related to birds^{11–13}. Thus, the proposed homology between the integumentary appendages of *Sinornithosaurus* and avian feathers is supported by observations of detailed, derived morphological similarities, and is strongly

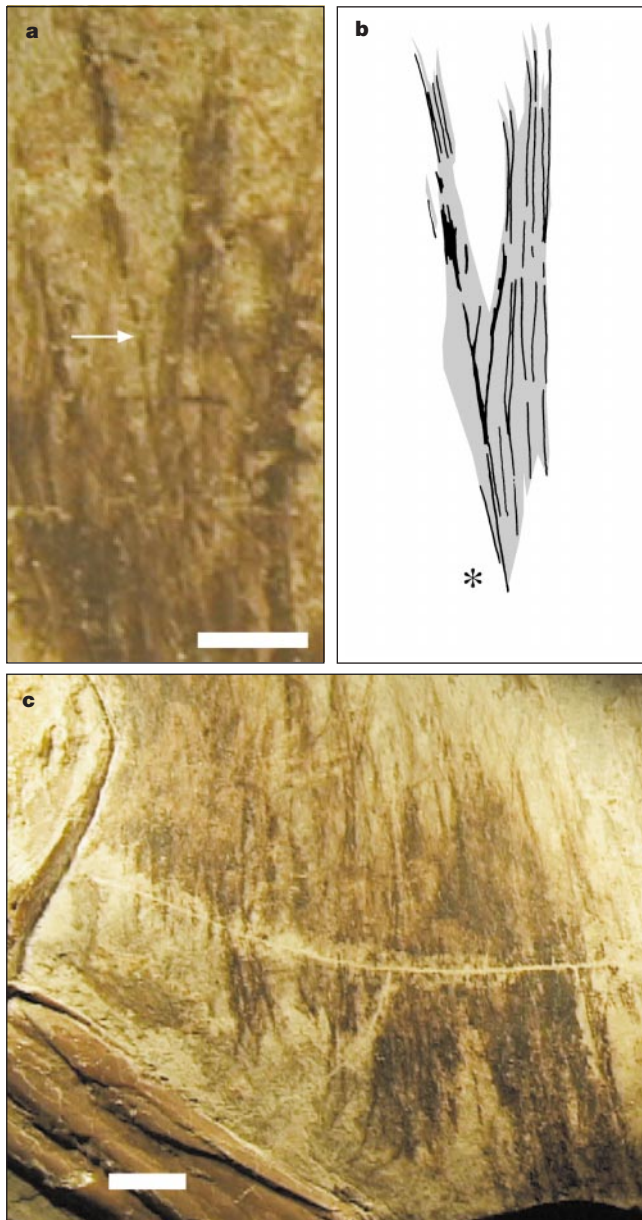


Figure 5 Filamentous integumental appendages of *S. millenii* indicating the presence of larger central filaments, or shafts, and secondary branched filaments. **a**, Dissociated appendages from near the left femur. Arrow indicates where a secondary filament joins at its base at an acute angle to a larger central filament. **b**, Illustrated reconstruction of the appendage showing the positions of the observed filaments (lines) and the inferred outline of the entire appendages (shading). Asterisk, the proximal end of the appendages. **c**, Filamentous integumental appendages preserved in nearly their original positions along the right ulna. Scale bars, 5 mm.

corroborated by independent phylogenetic data showing that these lineages are historically closely related.

The morphology of the plesiomorphic feathers of *Sinornithosaurus* is also congruent with the predictions of an independent, developmental model of the evolutionary origin of feathers¹⁴. This developmental model proposes that feathers originated with the evolution of the first cylindrical feather follicle, and that they subsequently evolved through a series of derived novelties in the developmental mechanisms within the follicle¹⁴. The model predicts a transition series of feather morphologies from the first hollow, cylindrical feather through all modern feather structural diversity¹⁴ (Fig 6). The two appendage morphologies observed

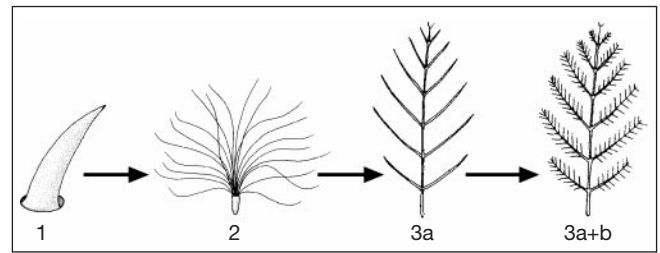


Figure 6 The predicted transition series of feather morphologies from an independent, developmental model of the evolutionary origin of feathers¹⁴. The model¹⁴ hypothesizes that the first feather originated with the first follicle—a cylindrical epidermal invagination around a papilla. Subsequent feather diversity evolved through a series of derived novelties in the developmental mechanisms within the follicle. The integumental appendages of *Sinornithosaurus* described here are congruent with Stages II and IIIa, whereas the integumental appendages of *Sinosauropteryx*² are congruent in morphology with Stage I. Stage I, an unbranched, hollow filament; Stage II, a tuft of barbs basally fused to a calamus; Stage IIIa, a feather with a rachis and serially fused barbs; Stage IIIa+b, a feather with rachis, barbs and barbules. Stage IV (not shown), a bipinnate feather with differentiated distal and proximal barbules and a closed pennaceous vane; Stage Va–f (not shown), additional modern feather diversities including asymmetrical flight feathers.

in *Sinornithosaurus* are exactly congruent with the Stage II and Stage IIIa morphologies predicted by this model (Fig. 6). Furthermore, the shorter, unbranched integumental appendages of *Sinosauropteryx*², a basal coelurosaur, are also congruent with the predicted Stage I feather morphology (Fig. 6). The morphologies and phylogenetic distributions of these theropod integumental appendages conform closely to the ancestral feather morphologies predicted by the developmental model of the origin of feathers¹⁴.

These results support the hypothesis that feathers evolved and initially diversified in terrestrial theropod dinosaurs before the origin of birds and flight. Furthermore, feathers evolved filamentous structure, basal branching, and a rachis with barbs before they evolved bipinnate structure, differentiated proximal and distal barbules, and the closed pennaceous vane that are required for aerodynamic function. Contrary to recent reports¹⁸, the hypothesis of the theropod origin of feathers is strongly corroborated by fossil observations and phylogenetic analyses. □

Received 2 August; accepted 19 December 2000.

- de Beer, G. *Archaeopteryx lithographica: A Study Based on the British Museum Specimen* (Trustees of the British Museum, London, 1954).
- Chen, P.-J., Dong, Z. M. & Zhen, S. N. An exceptionally well-preserved theropod dinosaur from the Yixian formation of China. *Nature* **391**, 147–152 (1998).
- Ji, Q., Currie, P. J., Norell, M. A. & Ji, S.-A. Two feathered dinosaurs from northeastern China. *Nature* **393**, 753–761 (1998).
- Xu, X., Tang, Z.-L. & Wang, X.-L. A therizinosauroid dinosaur with integumentary structures from China. *Nature* **399**, 350–354 (1999).
- Xu, X., Wang, X.-L. & Wu, X.-C. A dromaeosaurid dinosaur with a filamentous integument from the Yixian Formation of China. *Nature* **401**, 262–266 (1999).
- Xu, X., Zhou, Z. & Wang, X. The smallest known non-avian theropod dinosaur. *Nature* **408**, 705–708 (2000).
- Gibbons, A. Plucking the feathered dinosaur. *Science* **278**, 1229–1230 (1997).
- Feduccia, A. *The Origin and Evolution of Birds* 2nd edn (Yale Univ. Press, New Haven, 1999).
- Dalton, R. Feathers fly in Beijing. *Nature* **405**, 992 (2000).
- Jones, T. D., Farlow, J. O., Ruben, J. A., Henderson, D. M. & Hillenius, W. J. Cursoriality in bipedal archosaurs. *Nature* **406**, 716–718 (2000).
- Gauthier, J. A. Saurischian monophyly and the origin of birds. *Mem. California Acad. Sci.* **8**, 1–55 (1986).
- Holtz, T. R. The phylogenetic position of the Tyrannosauridae: implications for theropod systematics. *J. Paleontol.* **68**, 1100–1117 (1994).
- Sereno, P. The evolution of dinosaurs. *Science* **284**, 2137–2147 (1999).
- Prum, R. O. Development and evolutionary origin of feathers. *J. Exp. Zool. Mol. Dev. Evol.* **285**, 291–306 (1999).
- Lucas, A. M. & Stettenheim, P. R. *Avian Anatomy—Integument* (US Dept. Agriculture Handbook, Washington DC, 1972).
- Nowak, R. M. & Paradiso, J. L. *Walker's Mammals of the World* (Johns Hopkins Univ. Press, Baltimore, 1983).
- Grau, V. J. *Biología y Patología de la Chinchilla* (Ediciones OIKOS, Santiago, Chile, 1994).
- Jones, T. A. et al. Non-avian feathers in a late Triassic archosaur. *Science* **288**, 2202–2205 (2000).

Acknowledgements

We thank M. Chang and Y. Wang for assistance during the work; S.-h. Xie for preparation of the holotype specimen of *Sinornithosaurus*; and members of the Liaoxi expedition team of the IVPP, who participated in the excavation of the specimen. The paper benefited from discussions with C.-M. Chuong, P. Currie, A. Deane, L. Martin, B. Timm, G. Wagner, L. Witmer and F. Zhang; and was improved by comments from W. Boles, P. Stettenheim and L. Witmer. We acknowledge financial support from the Chinese Academy of Sciences, Chinese Natural Science Foundation, Chinese Special Funds for Major State Basic Research Projects (X.X., Z.Z.), and the United States National Science Foundation (R.O.P.). Order of authorship was assigned randomly.

Correspondence and requests for materials should be addressed to R.O.P. (e-mail: prum@ukans.edu).

Sexual swellings advertise female quality in wild baboons

Leah G. Domb*† & Mark Pagel‡

* Department of Anthropology, Harvard University, Cambridge, Massachusetts 02138, USA

‡ School of Animal and Microbial Sciences, University of Reading, Whiteknights, Reading RG6 6AJ, UK

The females of many Old World primate species produce prominent and conspicuous swellings of the perineal skin around the time of ovulation. These sexual swellings have been proposed to increase competition among males for females¹ or to increase the likelihood of a female getting fertilized, by signalling either a female's general reproductive status^{1–5}, or the timing of her ovulation⁶. Here we show that sexual swellings in wild baboons reliably advertise a female's reproductive value over her lifetime, in accordance with a theoretical model of honest signalling⁷. Females with larger swellings attained sexual maturity earlier, produced both more offspring and more surviving offspring per year than females with smaller swellings, and had a higher overall proportion of their offspring survive. Male baboons use the size of the sexual swelling to determine their mating effort, fighting more aggressively to consort females with larger swellings, and spending more time grooming these females. Our results document an unusual case of a sexually selected ornament in females, and show how males, by mating selectively on the basis of the size of the sexual swelling, increase their probability of mating with females more likely to produce surviving offspring.

The females of most Old World primate species with sexual swellings live in multi-male breeding systems, in which females typically mate with more than one male during oestrus^{1,2,4}. Theoretical models show that male–male competition for females is inevitable in these circumstances⁷, raising the issue of why females produce such exaggerated displays.

One of us (L.G.D.) investigated the function of sexual swellings in a population of wild olive baboons (*Papio cynocephalus anubis*) living in Gombe National Park, Tanzania. This population has been studied extensively since 1967 (refs 8–16). Olive baboons reside in troops of adult males and females, and their juvenile and infant offspring. Most males are immigrants as females remain in their natal territories, and the breeding system is multi-male such that each female typically mates with more than one male per oestrous cycle.

Twenty-nine parous females for which long-term demographic

data were available were followed in the wild for 13 months. Measurements of the size of females' sexual swellings and of the behaviour of males towards females were collected for 22 females in 44 oestrus cycles. All male behaviour and sexual swelling-size data are from the 5 days before swelling detumescence in females. For the same females, we extracted from long-term demographic records information on four variables related to fitness: age at first conception; the number of offspring produced per year; the number of surviving offspring produced per year; and the proportion of a female's offspring that survived (see Methods).

Figure 1 shows an example of a baboon sexual swelling in its fully swollen state. We characterized each female's swelling by measuring its length, width and depth. Neither width nor depth of swellings was consistently related to the four indicators of female reproductive value (Table 1), and so hereafter we report only associations with length ('swelling size').

Length of a female's swelling correlates with measures of female fitness (Table 1): females with larger swellings begin to reproduce at an earlier age ($P < 0.005$), produce a larger number of offspring per year ($P < 0.01$) and a larger number of surviving offspring per year ($P < 0.025$); in addition, a greater proportion of their offspring survive ($P < 0.065$). All P values we report are for two-tailed tests. A female's swelling size was related to her age ($r = -0.61$, $P < 0.01$; $n = 22$), but not to her social rank ($r = 0.087$, $n = 18$). We nevertheless controlled statistically for both variables (Table 1). These analyses show that the relationships between the measures of female reproductive value and swelling size remain significant, and in some cases are strengthened. When controlling for female age and rank, swelling size is strongly correlated with the proportion of a female's offspring that survive ($P < 0.025$, Table 1). Size of the oestrous swelling therefore seems to be an indicator of a female's lifetime reproductive value, independently of age and social rank within the troop.

An important component of lifetime reproductive value is the probability of conceiving in each cycle. In humans, early maturing girls have higher frequencies of ovulation than later maturing girls, and higher levels of ovarian steroids in each cycle¹⁷. Early maturation also has been found in humans to be associated with better reproductive performance throughout a female's life^{18–19}. Among females in our study, age of first conception, in addition to being associated with larger swelling size, is significantly associated with a higher number of both offspring produced per year ($r = -0.43$, $P < 0.05$; $n = 29$) and surviving offspring produced per year, relative to other troop members ($r = -0.55$, $P < 0.01$; $n = 29$). Larger swellings signal not only that the female started reproducing earlier, but also that she has a higher reproductive rate relative to other group members. In a mating system such as that of olive baboons in



Figure 1 A baboon sexual swelling showing the fully swollen state.

† Present address: 38 Sydenham Road, Bristol BS6 5SJ, U.K.

which males contribute little paternal care, the ability of a female to conceive early in life, to have less time between conceptions, and to produce and rear surviving offspring successfully on her own may be the features of greatest interest to males. If so, it will pay females to advertise these traits.

In contrast to the associations across females between swelling size and reproductive value, we found no differences in the size of a given female's swellings in conception versus non-conception cycles. We had paired data for six females on conception and the nearest non-conception cycle. Neither mean length (conception = 20.2 cm, non-conception = 19.7 cm; $P = 0.36$ matched-pairs t -test) nor mean depth (conception = 14.0 cm, non-conception depth = 12.8 cm; $P = 0.45$ matched-pairs t -test; note that depth is sum of two measures, see Methods) differed between these cycles. Our sample size for these tests is small, but other primate studies also show that the size of a swelling is not related either to the current probability of conception²⁰, or to hormonal profiles indicative of ovulation^{20–22}.

If swelling size advertises a female's lifetime reproductive value, we should expect males to respond by showing more interest in females with larger swellings. Male interest in a given female was measured by noting the average consorting rank of the male ('consort rank'), the rank and persistence of additional males following the consorting couple ('followers'), the relative amount of time the female was groomed by the consort male ('grooming'), and the amount of aggression ('aggression') that the consort male received from other males while acting as a consort (see Methods).

The four measures of interest are intercorrelated, in particular aggression, consort rank and followers (Table 2), indicating that females that attracted more interest tended to be consorted by higher ranking males, and had higher ranking and more persistent followers; in addition, the consorting male suffered more aggression from other males. The relationship between aggression and followers was particularly strong; this is expected as most aggression towards a consorting male is initiated by follower males. Grooming, on the other hand, seemed to be independent of the other measures of interest.

Table 3 shows that males have greater interest in females with larger swellings. Males consorting females with larger swellings suffered substantially more aggression ($P < 0.005$) from other males, as compared with males consorting females with smaller swellings, even after controlling for female age and a female's family rank. There is also some tendency for females with large swellings to receive more grooming ($P < 0.05$). Thus, to consort a female with a large swelling, the male must endure greater costs of aggression from competing males and spend more time grooming the female. It is interesting to note that swelling size was not significantly correlated with male consort rank or the follower score.

By comparison, male interest was not significantly related directly to the measures of female reproductive potential (Table 4). This is not wholly unexpected. As immigrants to the troop, males have limited knowledge of female fitness, and therefore must rely on the female's signal to advertise her fitness and to determine their mating effort: Tables 1 and 3 show that this is what males seem to do. A similar situation has been reported in humans: Kipsigi men

pay higher bride prices for women who reach menarche earlier, although the men have no knowledge of the dates of this maturational step in their prospective brides²³.

Together, our results show that sexual swellings advertise females' long-term reproductive value, and that males use the swellings to determine their mating effort. Males are more interested in and expend greater effort in consorting females whose large swellings advertise a higher reproductive rate, and whose offspring have a higher probability of survival. Males mating females with large swellings therefore gain the immediate, direct benefits of having a higher probability of producing an offspring in a given mating attempt. They also gain the longer term benefit of having offspring that are more likely to survive. If this higher probability of offspring survival arises from a genetic quality inherited from the females, then males also potentially derive indirect or 'good-genes' benefits from mating females with larger swellings. A female's swelling size and reproductive value may arise from her phenotypic condition, from a heritable genetic quality, or both (for example, see ref. 24).

The alternative explanation for our results derives from receiver-bias theories of sexual selection (for example, see ref. 25). In this view, male interest in females with larger swellings arises from a male perceptual bias for features of the females' swellings, indepen-

Table 2 Correlations of male-interest scores

	Aggression	Consort rank	Followers
Consort rank	0.40		
Followers	0.65*	0.38	
Grooming	0.17	0.31	0.24

Standardized male-interest scores are given. Thirteen females had male-interest and swelling-size data.

* $P < 0.025$.

Table 3 Correlations between swelling size and male-interest scores

Swelling measurement	Aggression	Consort rank	Followers	Grooming
Length	0.78†	0.07	0.03	0.58*
Length (age, rank)	0.82†	-0.02	0.12	0.38

'Length (age, rank)' is the partial correlation controlling for female age and family rank, thereby recording length's unique contribution to the male-interest variables.

Thirteen females had male-interest and swelling-size data.

* $P < 0.05$.

† $P < 0.005$.

Table 4 Correlations between female characteristics and female-fitness measures and male-interest scores

	Aggression	Consort rank	followers	grooming
Age first conception	-0.45	-0.25	-0.001	-0.61*
No. offspring per year	0.44	0.40	0.31	0.16
No. surviving offspring per year	0.12	-0.37	-0.21	-0.08
Proportion of offspring surviving	-0.01	-0.51	-0.28	-0.13
Female age	-0.18	-0.15	0.15	-0.64†
Female rank	0.03	-0.06	-0.09	-0.26

Thirteen females had male-interest and swelling-size data.

* $P < 0.05$.

† $P < 0.025$.

Table 1 Correlations between swelling sizes and female-fitness measurements

Swelling measurement	Age at first conception	No. of offspring per year	No. of surviving offspring per year	Proportion of offspring surviving	<i>n</i>
Length	-0.67§	0.55‡	0.53†	0.40*	22
Depth	-0.52†	0.30	0.38	0.29	20
Width	-0.29	0.10	0.27	0.36	21
Length (age, rank)	-0.61†	0.58†	0.68§	0.59†	18

'Length (age, rank)' is the partial correlation controlling for female age and family rank, thereby recording length's unique contribution to the female-fitness measures. Swelling-length data were available for 22 females, swelling-width data for 21 females, and swelling-depth data for 20 females. Family rank was known for 18 of these females.

* $P < 0.065$.

† $P < 0.025$.

‡ $P < 0.01$.

§ $P < 0.005$.

dently of female fitness. This seems an unlikely explanation for our results. Receiver-bias theories make no link between the size or nature of the sexual signal and fitness, and therefore do not account for the strong association that we find between swelling size and female reproductive value (Table 1). Similarly, the receiver-bias view cannot, on its own, explain why males should be willing to incur greater costs to mate females with larger swellings.

We therefore conclude that oestrous swellings advertise key aspects of female quality. Two features of the signalling system are likely to ensure its reliability. One is that swellings are almost certainly costly to produce. They are substantial in size: a female baboon's (*Papio c. ursinus*) body weight increases by about 14% when maximally swollen²⁶ and by about 25% in some groups of red colobus (*Colobus badius*)²⁷. Swellings thus may constitute a significant extra load for females when travelling as well as constrain movements and sitting postures. Furthermore, the nature of the sexual skin may attract insects or other parasites, and they are easily cut or torn²⁸, providing scope for infection. The second is that mating effort is costly in multi-male primates. Mate guarding significantly reduces foraging in male baboons²⁹. Males must also consort and groom females, and ward off other suitors. In our study, more attractive females incited greater male-male competition, and it is known that fights in baboons are potentially lethal: baboons are a highly dimorphic species in both body and canine size, indicating that intrasexual selection has been a consistent feature of baboon evolution. These mating costs place a premium on choosing the right mate. It seems that, among baboons, males who choose correctly will mate females more likely to produce surviving offspring. □

Methods

Baboons

L.G.D. followed 29 parous females from five troops of wild olive baboons at Gombe National Park in Tanzania between November 1994 and January 1996. Females' ages ranged from 6.6 to 20.2 yr, with an average of 12.4 yr. Observational data were collected on sexual swellings and on male behaviour towards a cycling female. To compare levels of male interest between females, only periods in which males had one female available to consort were considered. Of these, 13 females (with 19 cycles) had both sexual swelling size and male-interest data. Four of the thirteen females contributed data from between two and four oestrous cycles. The rest contributed one oestrous cycle.

Female data

We obtained data on swelling length for 22 females in 44 oestrous cycles, with a range of 1 to 4 cycles per female. Twenty-one of these females had swelling-width scores and twenty had depth scores. A video and meter-stick method was created to measure the length, width and depth of sexual swellings. Specifically, a female's swelling was filmed in a consistent fashion. Once the female moved off, a meter stick was placed at the location of the swelling by an assistant and then filmed from the same location that was used to film the swelling. Linear dimensions were calculated using the computerized image program, NIH Image (<http://rsbweb.nih.gov/ni-image/>).

'Length' was measured when viewed from behind from under the tail down to its lowest point between the legs. 'Depth' was the sum of two depth measurements made from the side, one under the tail, one at the ischial callosities, and 'width' was the sum of two measurements made from behind, the width of the swelling at the anus and the width of the swelling at the base of the ischial callosities.

'Family rank' refers to the rank of a female's matriline within the troop during the course of the study. Matriline ranks were available for 4 of 5 troops, and 18 females in total. At no time during the study did any of the matriline change rank. Matriline ranks were taken from monthly reports from the long-term social and demographic study at Gombe. Daily demographic data collected on the baboons at Gombe since 1967 allowed us to calculate a female's age, age at first conception, reproductive rate (number of offspring/reproductive years) and rate of surviving offspring (number of surviving offspring/reproductive years). Offspring survival is defined as surviving at least 2 yr.

Male-interest data

Behavioural data on males consorting individual females during their oestrus was recorded systematically in focal samples. Consort identity and consorting activity were also noted in the morning and afternoon even if no focal samples were obtained. Analyses were confined to the last 5 days of the period when females were consistently mate-guarded (consorted) by individual males. We recorded four male behaviours. 'Consort rank': the average rank of consorts that a female attracted (weighted by the proportion of time each male spent consorting), with the males' rank measured in terms of consorting success in the troop. 'Followers': following (or shadowing) is a characteristic male baboon

behaviour towards a consorting couple. The follower score measures both male rank and their persistence: it is the average consorting rank of the follower males (weighted by the proportion of time each male spent following), multiplied by the total number of follower minutes. 'Grooming': the relative time a male spends grooming divided by the total grooming and resting time. 'Aggression': aggression received by the consort male (threats, attacks) during the focal period. Each bout of aggression was given one point and was circumscribed by a 5-min period with no agonistic interactions. These four measures yielded a male-interest measure for each female and each oestrous cycle.

Statistical analyses

All analyses use females as independent units. Where we have data from several cycles on the same female, we used the average of her swelling-size and male-interest scores. In all analyses, control variables were removed in the standard way by partial correlation.

Received 7 August; accepted 28 November 2000.

- Clutton-Brock, T. & Harvey, P. in *Growing Points in Ethology* (eds Bateson, P. & Hinde, R.) 195–237 (Cambridge Univ. Press, Cambridge, 1976).
- Hrdy, S. & Whitten, P. in *Primate Societies* (eds Smuts, B., Cheney, D., Seyfarth, R., Wrangham, R. & Strusaker, T.) 370–384 (Univ. Chicago Press, Chicago, 1987). [AUTHOR: page numbers]
- Hrdy, S. in *The Evolution of Sex* (eds Bellig, R. & Stevens, G.) 101–136 (Harper and Row, New York, 1988).
- Sillen-Tullberg, B. & Moller, A. The relationship between concealed ovulation and mating systems in anthropoid primates: a phylogenetic analysis. *Am. Nat.* **141**, 1–25 (1993).
- Nunn, C. The evolution of exaggerated sexual swellings in primates and the graded signal hypothesis. *Anim. Behav.* **58**, 229–246 (1999).
- Hamilton, W. J. in *Primate Paternalism* (ed. Taub, D.) 309–335 (Van Nostrand Reinhold, New York, 1984).
- Pagel, M. The evolution of conspicuous oestrous advertisement in Old World monkeys. *Anim. Behav.* **47**, 1333–1341 (1994).
- Packer, C. Male transfer in olive baboons. *Nature* **255**, 219–220 (1975).
- Packer, C. Inter-troop transfer and inbreeding avoidance in *Papio anubis*. *Anim. Behav.* **27**, 1–36 (1979).
- Packer, C. Male dominance and reproductive activity in *Papio anubis*. *Anim. Behav.* **27**, 37–45 (1979).
- Nash, L. Troop fissioning in free-ranging baboons in the Gombe Stream National Park. *Am. J. Phys. Anthropol.* **44**, 63–77 (1976).
- Owens, N. The development of sociosexual behavior in free-living baboons, *Papio anubis*. *Behaviour* **57**, 241–259 (1976).
- Ransom, T. *Beach Troop of the Gombe* (Associated Univ. Press, New Jersey, 1981).
- Collins, D., Busse, C. & Goodall, J. in *Infanticide: A Comparative and Evolutionary Perspective* (eds Hausfater, G. & Hrdy, S.) 193–215 (Aldine, Hawthorne, New York, 1984).
- Packer, C., Collins, D. A., Sindimwo, A. & Goodall, J. Reproductive constraints on aggressive competition in female baboons. *Nature* **373**, 60–63 (1995).
- Packer, C., Tatar, M. & Collins, D. Reproductive cessation in female mammals. *Nature* **392**, 807–811 (1998).
- Ellison, P. Developmental influences on adult ovarian hormonal function. *Am. J. Hum. Biol.* **8**, 725–734 (1996).
- Ryder, N. & Westoff, C. *Reproduction in the United States* (Princeton Univ. Press, Princeton, 1971).
- Critescu, M. Differential fertility depending on the age of puberty. *J. Human Evol.* **4**, 521–524 (1975).
- Heistermann, M., Mohle, U., Vervaecke, H., van Elsacker, L. & Hodges, J. K. Application of urinary and fecal steroid measurements for monitoring ovarian function and pregnancy in the bonobo (*Pan paniscus*) and evaluation of perineal swelling patterns in relation to endocrine events. *Biol. Reprod.* **55**, 844–853 (1996).
- Thompson, J. et al. The Sulawesi crested black macaque (*Macaca nigra*) menstrual cycle: Changes in perineal tumescence and serum estradiol, progesterone, follicle-stimulating hormone, and luteinizing hormone levels. *Biol. Reprod.* **46**, 879–884 (1992).
- Whitten, P. & Russell, E. Information content of sexual swellings and fecal steroids in sooty mangabeys (*Cercocebus torquatus atys*). *Am. J. Primatol.* **40**, 67–82 (1996).
- Borgerhoff-Mulder, M. Early maturing Kipsigis women have higher reproductive success than late maturing women and cost more to marry. *Behav. Ecol. Sociobiol.* **24**, 145–153 (1989).
- David, P., Bjorksten, T., Fowler, K. & Pomiankowski, A. Condition-dependent signalling of genetic variation in stalk-eyed flies. *Nature* **406**, 186–188 (2000).
- Ryan, M. Sexual selection, sensory systems, and sensory exploitation. *Oxford Surv. Evol. Biol.* **5**, 156–195 (1990).
- Bielert, C. & Busse, C. Influence of ovarian hormones on the food intake and feeding of captive and wild female chacma baboons (*Papio ursinus*). *Physiol. Behav.* **30**, 103–111 (1983).
- Strusaker, T. *The Red Colobus Monkey* (Univ. Chicago Press, Chicago, 1975).
- Hausfater, G. Dominance and reproduction in baboons (*Papio cynocephalus*): a quantitative analysis. *Contrib. Primatol.* **7**, 1–150 (1975).
- Alberts, S., Altmann, J. & Wilson, M. Mate guarding constrains foraging activity of male baboons. *Anim. Behav.* **51**, 1269–1277 (1996).

Acknowledgements

We thank C. Packer and D. A. Collins for providing access to the reproductive histories of the females in this study; the Government of Tanzania for permission to carry out this study; K. Hawkes, C. Packer, R. Wrangham and D. A. Collins for comments on this work; and A. Sindimwo, M. Bwenda, I. Rukamata, M. Omari, S. Hamisi, R. Mkono, J. Bushingwa, F. Juma, and D. A. Collins for assistance in data collection at Gombe National Park. L.G.D.'s field work was supported by National Science Foundation and Wenner-Gren Foundation for Anthropological Research. Data collection at Gombe was supported by the Jane Goodall Institute. M.P. is supported by the Leverhulme Trust and the Natural Environment Research Council.

Correspondence and requests for materials should be addressed to L.G.D. (e-mail: lgdomb@aol.com).

Mice lacking the M3 muscarinic acetylcholine receptor are hypophagic and lean

Masahisa Yamada^{*,†}, Tsuyoshi Miyakawa^{‡,§}, Alokesh Duttaroy^{*}, Akihiro Yamanaka^{||}, Toru Moriguchi[¶], Ryosuke Makita[†], Masaharu Ogawa[†], Chieh J. Chou[#], Bing Xia^{*}, Jacqueline N. Crawley[‡], Christian C. Felder^{*,*}, Chu-Xia Deng^{**} & Jürgen Wess^{*}

^{*} Laboratory of Bioorganic Chemistry, # Diabetes Branch, and ^{**} Laboratory of Biochemistry and Metabolism, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, Maryland 20892, USA

[†] Laboratory for Cell Culture Development, RIKEN Brain Science Institute, Wako-shi, Saitama 351-0198, Japan

[‡] Section on Behavioral Pharmacology, Experimental Therapeutics Branch, National Institute of Mental Health, Bethesda, Maryland 20892, USA

^{||} Department of Pharmacology, Institute of Basic Medical Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8575, Japan

[¶] Laboratory of Membrane Biochemistry and Biophysics, National Institute on Alcohol Abuse and Alcoholism, Rockville, Maryland 20852, USA

^{*} Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, Indiana 46285, USA

Members of the muscarinic acetylcholine receptor family (M1–M5) have central roles in the regulation of many fundamental physiological functions^{1,2}. Identifying the specific receptor subtype(s) that mediate the diverse muscarinic actions of acetylcholine is of considerable therapeutic interest, but has proved difficult primarily because of a lack of subtype-selective ligands³. Here we show that mice deficient in the M3 muscarinic receptor ($M3R^{-/-}$ mice) display a significant decrease in food intake, reduced body weight and peripheral fat deposits, and very low levels of serum leptin and insulin. Paradoxically, hypothalamic messenger RNA levels of melanin-concentrating hormone (MCH), which are normally upregulated in fasted animals leading to an increase in food intake^{4,5}, are significantly reduced in $M3R^{-/-}$ mice. Intra-cerebroventricular injection studies show that an agouti-related peptide analogue lacked orexigenic (appetite-stimulating) activity in $M3R^{-/-}$ mice. However, $M3R^{-/-}$ mice remained responsive to the orexigenic effects of MCH. Our data indicate that there may be a cholinergic pathway that involves M3-receptor-mediated facilitation of food intake at a site downstream of the hypothalamic leptin/melanocortin system and upstream of the MCH system.

Peripheral M3 muscarinic acetylcholine receptors are thought to have a role in parasympathetic stimulation of smooth muscle contraction and glandular secretion^{1–3}. The M3 receptor subtype is also widely expressed throughout the central nervous system including the diencephalon⁶. However, the physiological roles of these central M3 receptors remain unknown. To explore the physiological functions of the M3 muscarinic receptor subtype, we used gene-targeting technology to generate M3-receptor-deficient mice ($M3R^{-/-}$ mice) (Fig. 1).

Homozygous $M3R^{-/-}$ mice were obtained at the expected mendelian frequency, and did not differ from their wild-type littermates in overall health, fertility and longevity. Histological examinations showed no abnormalities in brain, heart, liver, lung, spleen and pancreas (data not shown). The absence of functional M3 receptor protein in $M3R^{-/-}$ mice was confirmed by western blot and immunoprecipitation using M3 receptor-selective antibodies (Fig. 1c, d).

Immunoprecipitation also indicated that the M3 receptor is expressed at relatively low densities (20–50 fmol mg⁻¹) in all principal brain areas, including cerebral cortex and diencephalon (thalamus/hypothalamus) (Fig. 1d; and data not shown). Moreover, the lack of M3 receptors did not lead to compensatory changes in the expression levels of the remaining four muscarinic receptor subtypes (M1, M2, M4 and M5; Fig. 1d).

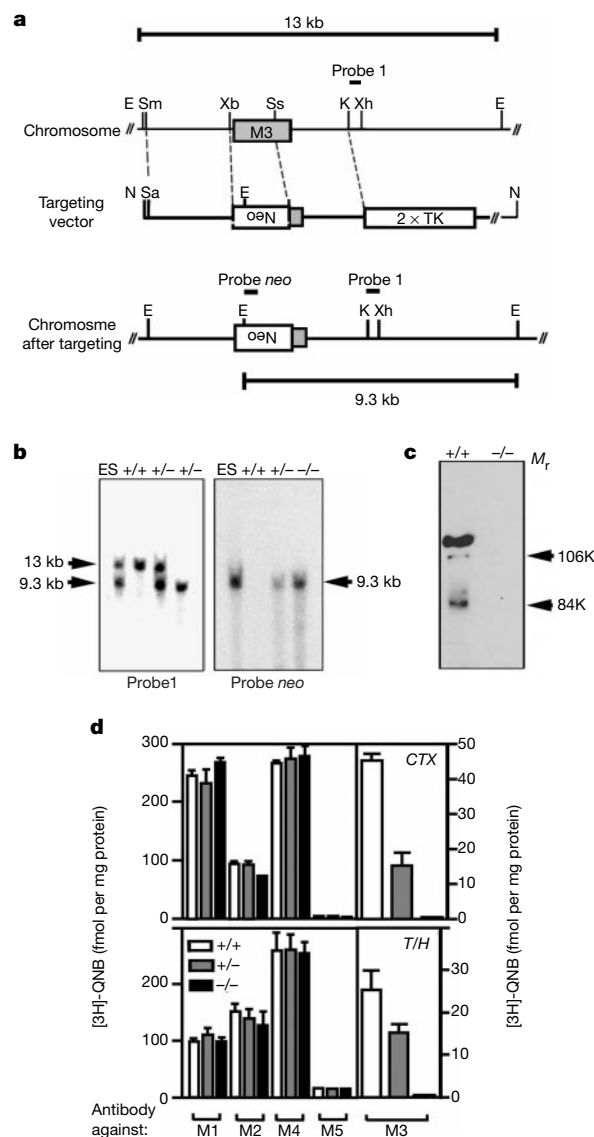


Figure 1 Targeted disruption of the mouse M3 muscarinic receptor gene. **a**, Structure of the wild-type allele, gene-targeting vector and mutated allele. The M3 receptor coding region is represented by a shaded box. Probes used for Southern analyses and sizes of the restriction fragments detected with these probes are indicated. E, EcoRV; K, KpnI; N, NotI; Sa, SalI; Sm, SmaI; Ss, Sse8337I; Xb, XbaI; Xh, XhoI. **b**, Southern blot analysis of both EcoRV-digested genomic DNA from a properly targeted ES cell clone and mouse tail DNA prepared from F₂ pups generated by inter-mating F₁ heterozygotes. The 13- and 9.3-kb bands represent the wild-type and mutant M3 receptor alleles, respectively. **c**, Western blot of mouse cerebral cortex membranes, using a polyclonal antibody directed against the carboxy terminus of the M3 receptor²⁸. The blot indicates the lack of the M3 receptor protein in $M3R^{-/-}$ mice. **d**, Immunoprecipitation of muscarinic receptor subtypes from mouse cerebral cortex (CTX) and the hypothalamus/thalamus (T/H) region. Twenty-week-old male mice of the indicated M3 genotypes were used. [3H]-QNB-labelled muscarinic receptors were solubilized and immunoprecipitated with M1–M5 receptor-specific antisera^{8,9}. Data are presented as mean ± s.d. (n = 4).

^{*} Present address: Department of Pharmacology & Center for Molecular Neuroscience, Vanderbilt University Medical Center, Nashville, Tennessee 37232-6600, USA.

During the first postnatal week, the body weight of $M3R^{-/-}$ mice was similar to that of their wild-type littermates (Fig. 2a, b). Starting at the end of postnatal week two (females) or three (males), however, $M3R^{-/-}$ mice displayed a pronounced decrease in body weight and reduced weight gain. At postnatal week 12, for example, male and female $M3R^{-/-}$ mice weighed about 22% less than their wild-type littermates (Fig. 2a, b). A similar percentage weight difference persisted throughout the entire observation period (>1 year). No significant reduction in body weight was observed with heterozygous $M3R^{+/-}$ mutant mice (data not shown). The body length (nose–anus) of 20-week-old male mice showed that linear growth was not significantly different in $M3R^{-/-}$ mice compared to wild-type littermates ($M3^{+/+}$, 9.87 ± 0.08 cm; $M3^{-/-}$, 9.71 ± 0.09 cm; mean $\pm$ s.e.m.; $n = 8$ for each group). Moreover, the pronounced reduction in body weight displayed by the $M3R^{-/-}$ mice was specific for loss of the M3 muscarinic receptor subtype, as this phenotype was not observed with mice lacking M1 (ref. 7), M2 (ref. 8), or M4 (ref. 9) muscarinic receptors.

$M3R^{-/-}$ mice (10- and 22-week-old males) had a 50–60% reduction in the weight of gonadal (epididymal) fat pads (Fig. 2g). As the weight of mouse gonadal fat pads closely correlates with total body fat mass¹⁰, reduced body fat content is predicted to be a main factor in the decreased body weight displayed by the $M3R^{-/-}$ mice.

To exclude the possibility that the observed mutant phenotype was caused by the mixed genetic background (C57BL/6J $\times$ 129SvEv) of the mice used for all experiments, we also examined isogenic mice that carried the M3 receptor mutation in a pure genetic background (129SvEv). Adult isogenic $M3R^{-/-}$ mice (14–24-week-old males; $n = 8$ –15 per group) did not differ significantly from their wild-type littermates in linear growth, but displayed reductions in body weight (by about 25%) and gonadal fat-pad mass (by about 60%) that were similar in magnitude to those found with mice of mixed

genetic background (data not shown).

Consistent with the observed lean phenotype, freely fed $M3R^{-/-}$ mice showed significantly (~6-fold) reduced levels of serum leptin (Table 1), a hormone that circulates at levels proportional to total body fat content^{11–13}. Moreover, serum triglyceride levels were found to be reduced by about 25% in $M3R^{-/-}$ mice (Table 1; $P < 0.05$). In contrast, the levels of many other serum parameters (total protein, total albumin, sodium, potassium, chloride, bicarbonate, calcium, alanine aminotransferase, aspartate aminotransferase, lactate dehydrogenase, alkaline phosphatase, creatine phosphokinase, urea nitrogen, creatinine, cholesterol and bilirubin) were not significantly affected by the lack of functional M3 receptors (Table 1; and data not shown).

To examine whether the reduction in body weight observed with $M3R^{-/-}$ mice was caused by altered feeding behaviour, we monitored daily food consumption (standard food pellets) over a 5-day period.

Table 1 Serum hormone and metabolite levels in $M3^{-/-}$ and wild-type mice

		$M3^{+/+}$	$M3^{-/-}$
Leptin	(ng ml ⁻¹)	13.50 ± 4.52	$2.14 \pm 0.18^*$
Insulin	(ng ml ⁻¹)	1.98 ± 0.62	$0.23 \pm 0.03^*$
Triiodothyronine (T3)	(ng l ⁻¹)	719 ± 62	765 ± 88
Thyroxine (T4)	(μg l ⁻¹)	25 ± 02	25 ± 01
IGF-I	(ng ml ⁻¹)	599 ± 48	516 ± 30
Corticosterone	(ng ml ⁻¹)	363 ± 41	476 ± 46
Triglycerides	(mg l ⁻¹)	$1,315 \pm 102$	$993 \pm 94^*$
Cholesterol	(mg l ⁻¹)	683 ± 71	557 ± 47
Total protein	(g l ⁻¹)	49.3 ± 3.0	42.5 ± 2.8
Albumin	(g l ⁻¹)	27.8 ± 0.8	2.88 ± 0.10
Glucose	(mg ml ⁻¹)	1.86 ± 0.23	1.63 ± 0.09

Blood from 18–22-week-old, freely fed male mice was collected from the aorta thoracica under anaesthesia between 10:00 and 12:00. All values are means $\pm$ s.e.m. ($n = 6$). * $P < 0.05$; Student's *t*-test.

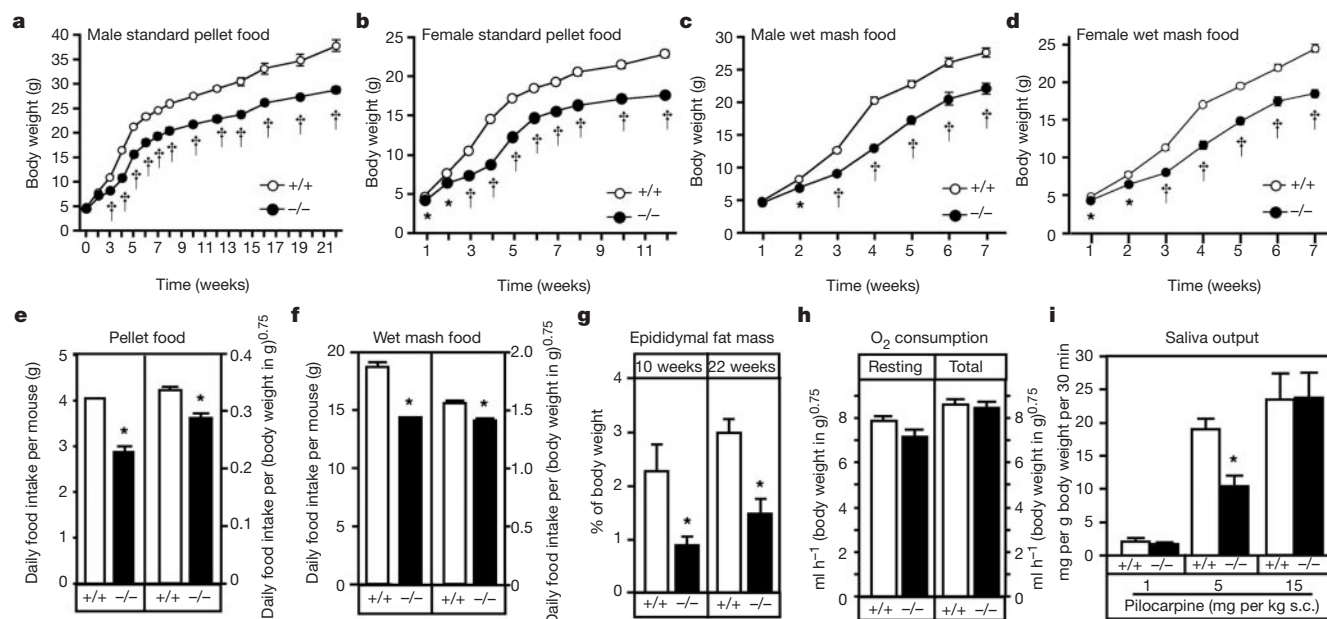


Figure 2 Body weight, food intake, mass of peripheral fat pads, metabolic rate and salivation response of wild-type and $M3R^{-/-}$ mice. **a, b**, Growth curves for male (**a**) and female (**b**) wild-type (open circles) and $M3R^{-/-}$ mice (filled circles) ($n = 17$ or 18 per group) fed with standard food pellets. **c, d**, Growth curves for male (**c**) and female (**d**) wild-type (open circles) and $M3R^{-/-}$ mice (filled circles) ($n = 12$ –19 per group) fed with wet mash food. Pups were weaned at the start of the third postnatal week. **e**, Daily consumption of standard food pellets by 12-week-old male mice ($n = 10$ per group) over 5 days. Data are expressed as food intake per mouse or are shown normalized to (body weight)^{0.75} (refs 19, 29). **f**, Daily consumption of wet mash food by 20-week-old male

mice ($n = 10$ per group). **g**, Percentage of epididymal fat-pad weight relative to total body weight (males, 10- or 22-weeks old; $n = 8$ per group). **h**, Resting and total oxygen consumption (metabolic rate), measured at room temperature in 12-week-old males ($n = 6$ per group; see Methods). Data were normalized to (body weight)^{0.75} (refs 19, 29). **i**, Muscarinic-receptor mediated salivation. Anaesthetized mice (24-week-old males; $n = 5$ or 6 per group) were injected (s.c.) with the indicated doses of the muscarinic agonist, pilocarpine. Saliva production was quantitated over 30 min by a filter paper method³⁰. Data are given as mean $\pm$ s.e.m.; * $P < 0.05$; † $P < 0.001$; Student's *t*-test.

We found that $M3R^{-/-}$ mice (12-week-old males) ingested about 30% fewer calories per day than their wild-type littermates (Fig. 2e), suggesting that the M3 receptor subtype has a role in regulating feeding behaviour.

M3 muscarinic receptors are thought to be involved in mediating cholinergic stimulation of salivary gland secretion¹⁻³. The reduced body weight of M3-receptor-deficient mice has been proposed to be caused by the absence of muscarinic receptor-mediated salivary flow¹⁴. We noted that the oral cavity of $M3R^{-/-}$ mice was wet, however, which suggests that basal salivary secretion was sufficient to permit normal feeding. Moreover, subcutaneous (s.c.) injection of different doses of the non-selective muscarinic agonist, pilocarpine, induced copious salivation in both wild-type and $M3R^{-/-}$ mice ($n = 5$ or 6 per group; 24-week-old males) (Fig. 2i). Whereas an intermediate dose (5 mg per kg (body weight)) of pilocarpine caused about a 50% reduction in saliva production, the lack of M3 receptors had no significant effect on the amount (Fig. 2i) and time course (data not shown) of salivation induced by 1 and 15 mg per kg of pilocarpine. These data suggest that other muscarinic receptor subtypes present in salivary gland tissue, such as M1 receptors^{3,15}, can maintain a robust salivation response in the $M3R^{-/-}$ mice.

We also carried out additional feeding experiments in which mice were offered a wet mash diet rather than dry food pellets. These studies showed that consumption of wet mash food instead of dry standard mouse chow had little effect on the reduction in body weight and food intake observed with the $M3R^{-/-}$ mice (Fig. 2c, d, f). Together, these observations argue against the idea that impaired salivation is responsible for the hypophagic phenotype displayed by the $M3R^{-/-}$ mice.

In the body periphery, M3 and M2 muscarinic receptors are co-expressed in gastrointestinal smooth muscle tissues where they are thought to promote gastric emptying and contractility of the gut¹⁶. Immunoprecipitation studies with stomach, duodenal, jejunal and ileal smooth muscle tissues showed that levels of the M2 receptor (which represents the predominant muscarinic receptor subtype in the gastrointestinal tract¹⁶) remained essentially unaffected by the lack of M3 receptors ($n = 4$; data not shown).

To exclude the possibility that food mal-absorption caused by impaired gastrointestinal motility contributed to the reduced body weight displayed by $M3R^{-/-}$ mice, we also carried out a charcoal transit test to monitor gastrointestinal smooth muscle activity *in vivo*. In this test, gastrointestinal transit of an intragastrically administered charcoal suspension is expressed as the ratio of the distance travelled by the charcoal during the 20-min period after

administration to the total length of the small intestine. The results of this analysis showed that the lack of M3 receptors had no significant effect on gastrointestinal motility *in vivo* ($M3^{+/+}$ mice, 0.56 ± 0.03 ; $M3^{-/-}$ mice, 0.53 ± 0.03 ; mean $\pm$ s.e.m.; $n = 6$ for each group). It is therefore unlikely that impaired gastrointestinal motor activity makes a significant contribution to the low body weight of $M3R^{-/-}$ mice.

Wild-type and $M3R^{-/-}$ mice (12–24-week-old males; $n = 18$ per group) were also subjected to behavioural and neurological tests¹⁷, including tasks that require attention and motivation (such as light/dark transition, accelerating rotarod, Morris water maze, contextual and cued fear conditioning, acoustic startle and pre-pulse inhibition). In all these tests, the $M3R^{-/-}$ mice performed equally well as their wild-type littermates (data not shown).

We next examined whether $M3R^{-/-}$ mice (freely fed 18–22-week-old males) showed altered serum levels of hormones involved in growth and the regulation of anabolic or catabolic pathways (time of blood collection: 10:00 to 12:00). Serum levels of insulin-like growth factor I (IGF-I), which is considered the main mediator of growth hormone action, remained unchanged in $M3R^{-/-}$ mice (Table 1). Similarly, serum levels of thyroid hormones (T3, T4) and corticosterone did not differ significantly among $M3R^{-/-}$ mice and their wild-type littermates (Table 1).

Notably, however, $M3R^{-/-}$ mice exhibited markedly reduced (9-fold) serum insulin levels (Table 1). Despite low circulating insulin levels, $M3R^{-/-}$ mice showed normal or slightly reduced blood (serum) glucose levels (Table 1) and did not become hyperglycaemic throughout the entire observation period (> 1 year; data not shown). Moreover, a glucose tolerance test (2 g per kg intraperitoneally (i.p.)) showed that $M3R^{-/-}$ mice were able to clear glucose from the blood at least as efficiently as wild-type mice (Fig. 3a). Similar results were obtained when the glucose load (2 g per kg) was administered orally rather than i.p. (data not shown), indicating that the $M3R^{-/-}$ mice maintain normal glucose tolerance despite marked hypoinsulinaemia. In fact, when mice were injected with a fixed dose of insulin (0.75 U per kg i.p.; insulin tolerance test), $M3R^{-/-}$ mice showed a much more pronounced and prolonged hypoglycaemia than their wild-type littermates, indicative of an increase in insulin sensitivity (Fig. 3b). The low serum insulin levels found in $M3R^{-/-}$ mice are probably due to the reduction in total body fat, as insulin, like leptin, circulates at levels proportional to body fat content¹³. We cannot, however, rule out the possibility that the absence of pancreatic M3 receptors, which may be involved in facilitating insulin release¹⁸, makes an additional contribution to the observed reduction in serum insulin levels.

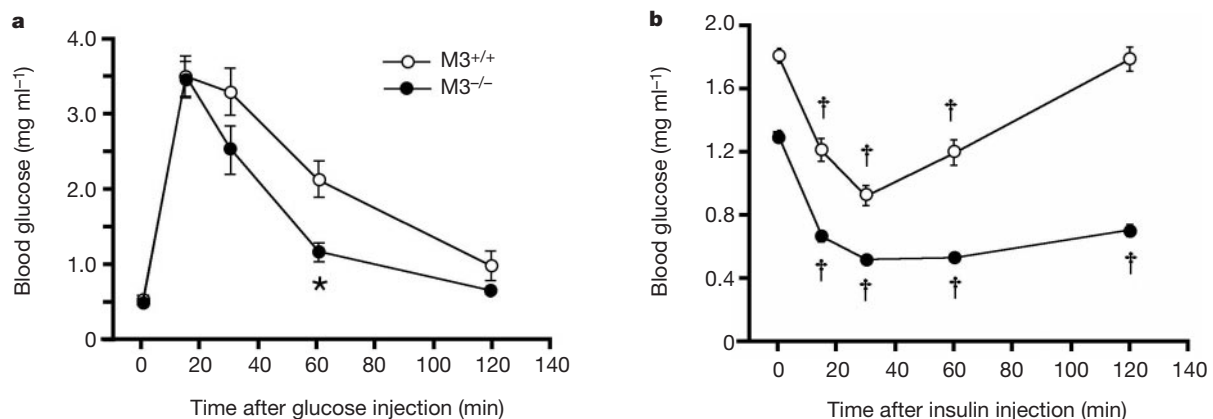


Figure 3 Glucose and insulin tolerance tests for wild-type and $M3R^{-/-}$ mice. **a**, Glucose tolerance test. Fasted 18-week-old male mice received an i.p. injection of glucose (2 g per kg (body weight)). **b**, Insulin tolerance test. Insulin was injected i.p. into 28-week-old,

freely fed male mice. Blood glucose levels were determined at the indicated times. Values are expressed as means $\pm$ s.e.m. ($n = 6$ per group). *, $P < 0.05$ (relative to wild type; **a**); †, $P < 0.001$ (relative to pre-injection values; **b**); Student's *t*-test.

To examine whether an increase in energy expenditure due to increased physical activity contributed to the lean phenotype displayed by the $M3R^{-/-}$ mice, we next carried out 48-h locomotor activity measurements using 17–20-week-old male mice. The lack of M3 receptors had no significant effect on horizontal locomotor activity, as indicated by the number of photobeam interruptions measured in 1-h intervals during the 24-h circadian cycle. Total numbers of photobeam breaks per 12 h were: $M3R^{-/-}$, light, $24,254 \pm 2,934$; $M3R^{+/+}$, light, $29,439 \pm 4,140$; $M3R^{-/-}$, dark, $59,011 \pm 8,011$; $M3R^{+/+}$, dark, $53,412 \pm 8,163$ (mean $\pm$ s.e.m.; $n = 13$ for $M3R^{-/-}$; $n = 12$ for $M3R^{+/+}$). $M3R^{-/-}$ mice (13-week-old males) showed a small but significant ($P < 0.05$) reduction in rectal temperature ($M3R^{+/+}$, 36.1 ± 0.1 °C; $M3R^{-/-}$, 35.5 ± 0.1 °C; means $\pm$ s.e.m.; $n = 20$ for $M3R^{-/-}$ and $n = 19$ for $M3R^{+/+}$).

Resting and total oxygen consumption (metabolic rate) of $M3R^{-/-}$ mice were not significantly different from the corresponding values determined with wild-type littermates (12-week-old males; $n = 6$ per group; Fig. 2h). Similarly, the respiratory exchange ratio (the ratio of carbon dioxide produced to oxygen consumed) and the degree of stimulation in metabolic rate in response to the β_3 -adrenergic receptor agonist, CL316243 (1 mg per kg i.p.), remained unaffected by the lack of M3 receptors (data not shown). These observations exclude the possibility that increases in exploratory locomotor activity, metabolic rate and/or energy consumption make a big contribution to the reduced body weight displayed by the $M3R^{-/-}$ mice.

Together, our data suggest that the low body weight of $M3R^{-/-}$ mice is attributable primarily to a reduction in food intake. Energy homeostasis and food intake are under strict control by regulatory centres in the hypothalamus^{11–13}. Radioligand binding studies with a saturating concentration (2 nM) of a non-selective muscarinic antagonist, [³H]quinuclidinyl benzilate (QNB), showed that muscarinic receptors were expressed at high density in the hypothalamus of wild-type mice (maximum number of binding sites ($B_{\max}$) = 2084 ± 157 fmol mg⁻¹; $n = 4$). Hypothalamic muscarinic receptor density was found to be reduced by about 50% ($P < 0.001$) in $M3R^{-/-}$ mice ($B_{\max} = 1087 \pm 169$ fmol mg⁻¹; $n = 4$), indicating that the M3 receptor subtype is abundantly expressed in this brain area. Moreover, immunohistochemistry showed that the M3 muscarinic receptor is expressed in different areas of the hypothalamus, including the arcuate nucleus and lateral hypothalamus (Fig. 4c).

Physiologically, reduced leptin and insulin levels are sensed by neurons in the arcuate nucleus of the hypothalamus, resulting in the activation of hypothalamic signalling pathways that lead to the stimulation of food intake and the suppression of energy consumption^{11–13}. One possible explanation for the failure of $M3R^{-/-}$ mice to respond properly to low circulating insulin and leptin levels with an increase in food intake is that the absence of M3 receptors interferes with appetite-regulating signalling pathways downstream of the primary leptin and insulin targets (receptors) in the arcuate nucleus. We therefore measured the expression levels of several

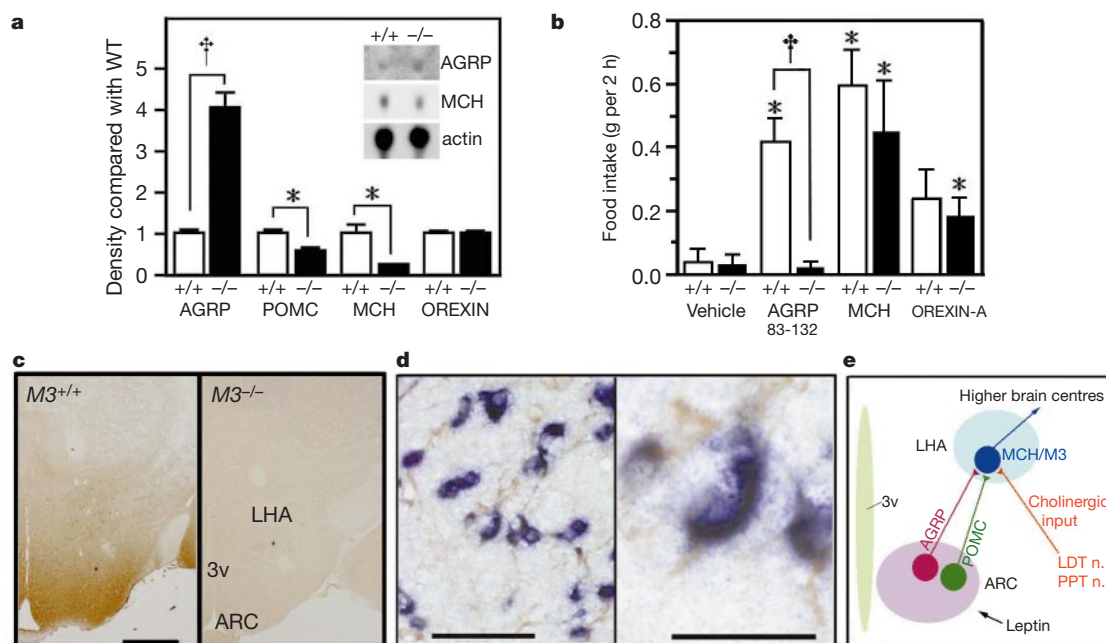


Figure 4 Role of M3 muscarinic receptors in neuropeptide-mediated feeding behaviour. **a**, Expression of hypothalamic neuropeptide mRNAs in wild-type and $M3R^{-/-}$ mice. Total RNA was isolated from hypothalami of 20-week-old, freely fed male mice and subjected to semi-quantitative RT-PCR analysis using primers specific for the AGRP, POMC, MCH, prepro-orexin and β -actin (internal standard) genes. For each sample, the ratio of the integrated optical density of the neuropeptide band to that of the internal standard was calculated. Neuropeptide expression levels of $M3R^{-/-}$ mice are expressed relative to those of wild-type mice (means $\pm$ s.e.m.; $n = 5$ per group). Hypothalamic levels of AGRP and MCH mRNAs were also studied by northern blot (insert), using 20 μ g of total RNA (pooled from the hypothalami of four mice per genotype). *, $P < 0.05$, †, $P < 0.001$; Student's t -test. **b**, Feeding response to intra-cerebroventricular administration of orexigenic neuropeptides. AGRP_{83–132} (0.9 nmol), MCH (0.2 nmol), or orexin-A (0.2 nmol) was given to 20-week-old male mice. Consumption of standard pellet food was measured for 2 h after injections. Data are expressed as mean $\pm$ s.e.m. ($n = 5$ or 6 per group), and were

analysed by ANOVA followed by Student's t -test (*, $P < 0.05$, relative to vehicle; †, $P < 0.001$). **c**, Expression of M3 receptors in the mouse hypothalamus studied by immunohistochemistry using a polyclonal antibody directed against the C terminus of the M3 receptor²⁸. M3 receptor immunoreactivity was observed in several regions of the hypothalamus, including the arcuate nucleus (ARC) and the lateral hypothalamus (LHA). Scale bar, 650 μ m. **d**, Presence of M3 receptor protein (brown) in neurons of the lateral hypothalamus expressing MCH mRNA (blue). Scale bars, 65 μ m (left); 25 μ m (right). **e**, Co-expression of MCH and M3 muscarinic receptors in neurons of the lateral hypothalamus (LHA), a secondary feeding centre. MCH neurons receive synaptic input from both the AGRP/POMC system of the arcuate nucleus (ARC)^{22–24}, a primary hypothalamic feeding centre, and cholinergic neurons of the laterodorsal tegment (LDT n.) and pedunculopontine tegment nuclei (PPT n.)²⁷. Our data indicate that M3 receptors may be required for the proper responsiveness of MCH neurons to cholinergic stimulation and input from the AGRP/POMC system.

hypothalamic neuropeptides (proopiomelanocortin (POMC), agouti-related peptide (AGRP), prepro-orexin, and melanin-concentrating hormone (MCH)) that are considered critical regulators of feeding and energy balance and act downstream of the leptin system^{11–13}.

Whereas POMC and AGRP-containing neurons are present in the arcuate nucleus and are the primary leptin targets^{11–13}, MCH²⁰ and the orexins²¹ are synthesized by second-order lateral-hypothalamus neurons that receive innervation from POMC and AGRP neurons^{22–24}. Semi-quantitative polymerase chain reaction with reverse transcription (RT–PCR) and northern blotting showed that hypothalamic AGRP mRNA levels were significantly increased, whereas hypothalamic POMC mRNA expression was reduced markedly in *M3R*^{−/−} mice (Fig. 4a). Consistent with the observation that *M3R*^{−/−} mice are hypophagic, this pattern (a low POMC:AGRP ratio) is characteristically seen in fasted mice or under conditions of leptin deficiency, and serves to stimulate food intake by reducing the activity of hypothalamic melanocortin receptors^{11–13}. Whereas prepro-orexin mRNA levels remained unaffected by the loss of M3 receptors, hypothalamic MCH expression was significantly reduced in *M3R*^{−/−} mice (Fig. 4a). MCH has a central role in the stimulation of food intake and the regulation of body weight^{4,5,11–13}. MCH expression is increased by reduced food intake and leptin deficiency in wild-type animals^{4,5,11–13}, and MCH-deficient mice are hypophagic and lean⁵. Moreover, impaired MCH signalling may contribute to the pathogenesis of certain forms of anorexia²⁵.

To examine the mechanism by which the lack of M3 receptors interferes with neuropeptide-dependent feeding behaviour, *M3R*^{−/−} and wild-type mice received intracerebroventricular infusions of AGRP_{83–132} (an AGRP analogue)²⁶, MCH and orexin-A. Interestingly, central administration of the melanocortin-4 receptor antagonist, AGRP_{83–132}, caused a robust stimulation of food intake in wild-type animals but had no effect on feeding behaviour in *M3R*^{−/−} mice (Fig. 4b). By contrast, the orexigenic effects of MCH and orexin-A remained intact in *M3R*^{−/−} mice (Fig. 4b), suggesting that reduced MCH expression may be a main factor in the hypophagia displayed by *M3R*^{−/−} mice.

In situ hybridization/immunohistochemistry double-labelling studies showed that most MCH mRNA-containing neurons (>90%) located in the lateral hypothalamus co-express M3 muscarinic receptors (Fig. 4d). M3 receptor protein was localized to both MCH cell bodies and dendritic processes. As hypothalamic MCH neurons receive abundant ascending cholinergic projections from the pedunculopontine and laterodorsal tegmental nuclei, and activation of hypothalamic muscarinic receptors leads to a robust increase in MCH mRNA expression²⁷, our findings are consistent with the concept that hypothalamic MCH expression is under stimulatory cholinergic control mediated by M3 muscarinic receptors.

In conclusion, our data indicate that there may be an M3-receptor-dependent cholinergic pathway that regulates appetite at a site downstream of the immediate hypothalamic leptin targets and upstream of the MCH system (Fig. 4e). Decreased MCH expression and reduced responsiveness of secondary hypothalamic feeding centres to the orexigenic effects of AGRP are predicted to represent two main factors responsible for the hypophagia phenotype displayed by the *M3R*^{−/−} mice (Fig. 4e). These findings may offer new therapeutic perspectives for the treatment of obesity. □

Methods

Generation of M3 receptor mutant mice

We isolated a murine M3 muscarinic receptor genomic clone from a 129Sv/J mouse genomic library (Genome Systems). The final targeting vector contained two copies of the herpes simplex virus thymidine kinase gene (*TK*) and a PGK–neomycin resistance cassette (*neo*), which replaced a 1.6-kilobase (kb) *Xba*I/*Sse*8337I genomic fragment that included the translation start site (Fig. 1a). The targeting vector was linearized and introduced into TC1 (129SvEv) embryonic stem (ES) cells by electroporation⁹. G418- and gancyclovir-

resistant clones were screened for the occurrence of homologous recombination by Southern blotting hybridization. M3 receptor mutant mice were generated essentially as described⁹. Unless stated otherwise, all experiments were carried out with littermates of the F₂ or F₃ generation (C57BL/6J × 129SvEv hybrids).

Western blotting and immunoprecipitation

Western blotting was carried out with membranes prepared from mouse cerebral cortex essentially as described²⁸, using an M3-receptor-specific rabbit polyclonal antibody. For immunoprecipitation, membrane preparations were incubated with 2 nM [³H]quinuclidinyl benzilate (QNB). [³H]QNB-labelled muscarinic receptors were solubilized with 1% digitonin and immunoprecipitated by M1–M5 receptor-specific antisera essentially as described^{8,9}.

Physiological measurements

Unless stated otherwise, all experiments were carried out between 9:00 and 12:00 using male littermates that were 10–22-weeks old. Mice were housed in a temperature- and humidity-controlled vivarium which was kept on a 12 h dark/light cycle (lights on 6:00, lights off 18:00). For food intake measurements, wild-type and *M3R*^{−/−} mutant mice were housed individually for 4 days before the start of the experiment. The consumption of standard dry rodent food pellets (Zeigler, Gardners, PA) or wet mash food (prepared by mixing powder made from CA-1 (CLEA) pellets (3.45 kcal g^{−1}) with twice the weight of sterilized water) were then measured for 5 consecutive days. Rectal temperature was measured at 10:00 with a rectal probe (Thermalert, Physitemp, Clifton, NJ).

For serum chemistry and hormone measurements, blood was collected from freely fed mice between 10:00 and 12:00. Serum hormone concentrations were measured by radioimmunoassay: insulin and leptin (Linco), T3 and T4 (Diagnostic Products), IGF-I (DSL) and corticosterone (ICN). Blood glucose levels were measured by glucose strips (ExacTech RSG glucose meter; Medisense). Serum chemistry measurements were carried out by ANI-LYTICS (Gaithersburg, MD) using a Hitachi 717 Chemistry Analyzer. For glucose tolerance tests, mice were fasted for at least 8 h and then given glucose (2 g per kg) through i.p. injection or oral gavage. For insulin sensitivity tests, human insulin (0.75 U per kg) (Eli Lilly) was administered intraperitoneally to freely fed mice. Blood samples were taken through a retro-orbital sinus puncture.

Oxygen consumption and carbon dioxide production were measured by using a four-chamber Oxymax system (Columbus Instruments) with one mouse per chamber, essentially as described¹⁹. To perform charcoal transit tests, mice that had been fasted for 24 h received intragastrically (by oral gavage) 0.2 ml of a suspension of 2% activated charcoal in 1.2% gum arabic. Mice were killed 20 min later, and gastrointestinal transit was expressed as the distance travelled by the charcoal per total length of the small intestine. For salivation studies, we anaesthetized mice with pentobarbital (50 mg per kg i.p.). Salivation responses to pilocarpine (1, 5 and 15 mg per kg subcutaneous (s.c.)) were quantitated in 5-min intervals over a 30-min observation period by a filter paper method³⁰.

Home cage locomotor activity

Home cage locomotor activity was quantitated using a Digiscan open-field arena (Accuscan Electronics). Horizontal activity of individually housed mice was measured as number of photobeam interruptions during 1-h intervals over a 48-h observation period.

RT–PCR

Total hypothalamic RNA (~2 µg) was reverse transcribed using oligo-dT primers and AMV reverse transcriptase (Promega). We carried out PCR amplifications with the following primer pairs: AGRP (GenBank accession number NM007427), 5′-TGACTGC AATGTTGCTGAGTTGTG-3′ and 5′-CTAGGTGCGACTACAGAGGTTTCGAG-3′ (product size: 392 base pairs (bp)); POMC (NM008895), 5′-TTTCTGGCAACGGAG ATGA-3′ and 5′-CCACCGTAACGCTTGTCCTT-3′ (449 bp); MCH, 5′-CAGCTTCC AAGTCCATAAGG-3′ and 5′-ACTCTTCCCAGCATACACCT-3′ (498 bp; ref. 10); prepro-orexin (NM011025), 5′-AGCCTCTGCCGACTGCTGT-3′ and 5′-ATGGTC AGGATGCCAGCTGC-3′ (191 bp); β-actin, 5′-ACCCACACTGTGCCATCTCA-3′ and 5′-CTAGAAGCATTTGCGGTGGA-3′ (641 bp). We quantified PCR products during the exponential phase of amplification (data available on request). For northern blotting, we used P³²-labelled DNA probes corresponding in sequence to the RT–PCR-amplified regions of the mouse AGRP and MCH mRNAs.

Intracerebroventricular infusions

Male mice (20-weeks old) were implanted with a guide cannula into the left lateral ventricle (0.3-mm posterior and 3.0-mm ventral to bregma) as described²¹. After a 7-day recovery period, peptides were dissolved in 2 µl of 0.9% saline and administered through the guide cannula at 2 h into the light cycle. Food intake was measured during the 2 h after injections. Mice had at least 48 h between treatments to regain normal feeding patterns. AGRP_{83–132} (human) and MCH were from Phoenix Pharmaceuticals; orexin-A (human) was obtained from the Peptide Institute.

Immunohistochemistry and *in situ* hybridization

Sections prepared from mouse hypothalami were processed for M3 receptor immunohistochemistry or for combined *in situ* hybridization/immunohistochemistry as described²². MCH mRNA was detected by the use of a digoxigenin-labelled MCH

ribochrome (blue reaction product)²². We subsequently processed sections for M3 receptor immunohistochemistry using an M3-receptor-specific rabbit polyclonal antibody (diluted 1:50)²⁸. Bound antibody was visualized by using an avidin/biotin/peroxidase method (brown reaction product)²².

Received 1 November; accepted 11 December 2000.

1. Brown, J. H. & Taylor, P. in *The Pharmacological Basis of Therapeutics* 9th edn (eds Hardman, J. G. et al.) 141–160 (McGraw-Hill, New York, 1996).
2. Wess, J., Buhl, T., Lambrecht, G. & Mutschler, E. in *Comprehensive Medicinal Chemistry*, Vol. 3 (ed. Emmett, J. C.) 423–491 (Pergamon, Oxford, 1990).
3. Caulfield, M. P. Muscarinic receptors—characterization, coupling and function. *Pharmacol. Ther.* **58**, 319–379 (1993).
4. Qu, D. et al. A role for melanin-concentrating hormone in the central regulation of feeding behaviour. *Nature* **380**, 243–247 (1996).
5. Shimada, M., Tritos, N. A., Lowell, B. B., Flier, J. S. & Maratos-Flier, E. Mice lacking melanin-concentrating hormone are hypophagic and lean. *Nature* **396**, 670–674 (1998).
6. Levey, A. I., Edmunds, S. M., Heilman, C. J., Desmond, T. J. & Frey, K. A. Localization of muscarinic M3 receptor protein and M3 receptor binding in rat brain. *Neuroscience* **63**, 207–221 (1994).
7. Hamilton, S. E. et al. Disruption of the m1 receptor gene ablates muscarinic receptor-dependent M current regulation and seizure activity in mice. *Proc. Natl Acad. Sci. USA* **94**, 13311–13316 (1997).
8. Gomez, J. et al. Pronounced pharmacologic deficits in M2 muscarinic acetylcholine receptor knockout mice. *Proc. Natl Acad. Sci. USA* **96**, 1692–1697 (1999).
9. Gomez, J. et al. Enhancement of D1 dopamine receptor-mediated locomotor stimulation in M4 muscarinic acetylcholine receptor knockout mice. *Proc. Natl Acad. Sci. USA* **96**, 10483–10488 (1999).
10. Rogers, P. & Webb, G. P. Estimation of body fat in normal and obese mice. *Br. J. Nutr.* **43**, 83–86 (1980).
11. Flier, J. & Maratos-Flier, E. Obesity and hypothalamus: Novel peptides for new pathways. *Cell* **92**, 437–440 (1998).
12. Elmquist, J. K., Elias, C. F. & Saper, C. B. From lesions to leptin: hypothalamic control of food intake and body weight. *Neuron* **22**, 221–232 (1999).
13. Schwartz, M. W., Woods, S. C., Porte, D. Jr, Seeley, R. J. & Baskin, D. G. Central nervous system control of food intake. *Nature* **404**, 661–671 (2000).
14. Matsui, M. et al. Multiple functional defects in peripheral autonomic organs in mice lacking muscarinic acetylcholine receptor gene for the M3 subtype. *Proc. Natl Acad. Sci. USA* **97**, 9579–9584 (2000).
15. Levey, A. I. Immunological localization of m1–m5 muscarinic acetylcholine receptors in peripheral tissues and brain. *Life Sci.* **52**, 441–448 (1993).
16. Egle, R. M., Hegde, S. S. & Watson, N. Muscarinic receptor subtypes and smooth muscle function. *Pharmacol. Rev.* **48**, 531–565 (1996).
17. Crawley, J. N. & Paylor, R. A proposed test battery and constellations of specific behavioral paradigms to investigate the behavioral phenotypes of transgenic and knockout mice. *Horm. Behav.* **31**, 197–211 (1997).
18. Boscher, A. C. et al. Oxotremorine-m potentiation of glucose-induced insulin release from rat islets involves M3 muscarinic receptors. *Am. J. Physiol.* **268**, E336–E342 (1995).
19. Gong, D. W. et al. Lack of obesity and normal response to fasting and thyroid hormone in mice lacking uncoupling protein-3. *J. Biol. Chem.* **275**, 16251–16257 (2000).
20. Bittencourt, J. C. et al. The melanin-concentrating hormone system of the rat brain: an immuno- and hybridization histochemical characterization. *J. Comp. Neurol.* **319**, 218–245 (1992).
21. Sakurai, T. et al. Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* **92**, 573–585 (1998).
22. Elias, C. F. et al. Chemically defined projections linking the mediobasal hypothalamus and the lateral hypothalamic area. *J. Comp. Neurol.* **402**, 442–459 (1998).
23. Broberger, C., De Lecea, L., Sutcliffe, J. G. & Hökfelt, T. Hypocretin/orexin- and melanin-concentrating hormone-expressing cells form distinct populations in the rodent lateral hypothalamus: relationship to the neuropeptide Y and agouti gene-related protein systems. *J. Comp. Neurol.* **402**, 460–474 (1998).
24. Elias, C. F. et al. Leptin differentially regulates NPY and POMC neurons projecting to the lateral hypothalamic area. *Neuron* **23**, 775–786 (1999).
25. Mystkowski, P. et al. Hypothalamic melanin-concentrating hormone and estrogen-induced weight loss. *J. Neurosci.* **20**, 8637–8642 (2000).
26. Rossi, M. et al. A C-terminal fragment of Agouti-related protein increases feeding and antagonizes the effect of alpha-melanocyte stimulating hormone in vivo. *Endocrinology* **139**, 4428–4431 (1998).
27. Bayer, L., Risold, P. Y., Griffond, B. & Fellmann, D. Rat diencephalic neurons producing melanin-concentrating hormone are influenced by ascending cholinergic projections. *Neuroscience* **91**, 1087–1101 (1999).
28. Zeng, F., Hopp, A., Soldner, A. & Wess, J. Use of a disulfide cross-linking strategy to study muscarinic receptor structure and mechanisms of activation. *J. Biol. Chem.* **274**, 16629–16640 (1999).
29. Kleiber, M. in *The Fire of Life* 2nd edn 179–222 (R. E. Krieger, Huntington, New York, 1975).
30. Parkes, M. W. & Parks, J. C. Supersensitivity of salivation in response to pilocarpine after withdrawal of chronically administered hyoscine in the mouse. *Br. J. Pharmacol.* **46**, 315–323 (1972).

Acknowledgements

This research was supported by the JSPS Research Fellowship Program (M.Y.) and through a CRADA between the NIDDK and the Eli Lilly Research Laboratories (B.X.). We thank M. L. Reitman, H. Kodama, H. Kanki and T. Sakurai for advice and discussions; J. Gan, N. Tsujino and C. Li for technical assistance; R. A. Kesterson for critical reading of the manuscript; and A. M. Spiegel and I. W. Levin for support of this work.

Correspondence and requests for materials should be addressed to J.W. (e-mail: jwess@helix.nih.gov).

NO-independent regulatory site on soluble guanylate cyclase

Johannes-Peter Stasch*, Eva Maria Becker*, Cristina Alonso-Alija*, Heiner Apeler*, Klaus Dembowski*, Achim Feuer*, Rupert Gerzer†, Torsten Minuth*, Elisabeth Perzborn*, Ulrich Pleiß*, Henning Schröder†, Werner Schroeder*, Elke Stahl*, Wolfram Steinke*, Alexander Straub* & Matthias Schramm*

* Pharma Research Center, Bayer AG, Aprather Wey 18a, D-42096 Wuppertal, Germany

† Martin Luther University, School of Pharmacy, Wolfgang-Langenbeck-Strasse 4, D-06099 Halle, Germany

‡ DLR, Institute of Aerospace Medicine, Linder Höhe, D-51147 Köln, Germany

Nitric oxide (NO) is a widespread, potent, biological mediator that has many physiological and pathophysiological roles¹. Research in the field of NO appears to have followed a straightforward path, and the findings have been progressive: NO and cyclic GMP are involved in vasodilatation; glycerol trinitrate relaxes vascular smooth muscles by bioconversion to NO; mammalian cells synthesize NO; and last, NO mediates vasodilatation by stimulating the soluble guanylate cyclase (sGC), a heterodimeric (α/β) haem protein that converts GTP to cGMP^{2–4}. Here we report the discovery of a regulatory site on sGC. Using photoaffinity labelling, we have identified the cysteine 238 and cysteine 243 region in the α_1 -subunit of sGC as the target for a new type of sGC stimulator. Moreover, we present a pyrazolopyridine, BAY 41-2272, that potently stimulates sGC through this site by a mechanism that is independent of NO. This results in antiplatelet activity, a strong decrease in blood pressure and an increase in survival in a low-NO rat model of hypertension, and as such may offer an approach for treating cardiovascular diseases.

In the search for NO-independent sGC stimulators, we examined around two thousand newly synthesized substances⁵. BAY 41-2272 is a representative (Fig. 1a) and potent stimulator that, at concentrations from 0.1 nM to 100 μ M, stimulates recombinant sGC (Fig. 1b). This stimulation was not blocked by high concentrations (65 μ M) of the NO scavengers oxyhaemoglobin and 2-phenyl-4,4,5,5-tetramethylimidazole-1-oxyl-3-oxide (PTIO). In combination, BAY 41-2272 and the NO donor 2-(N,N-diethylamino)-diazendate-2-oxide (DEA)/NO potentiate over a wide range of concentrations. At the highest concentrations of DEA/NO (1 μ M)

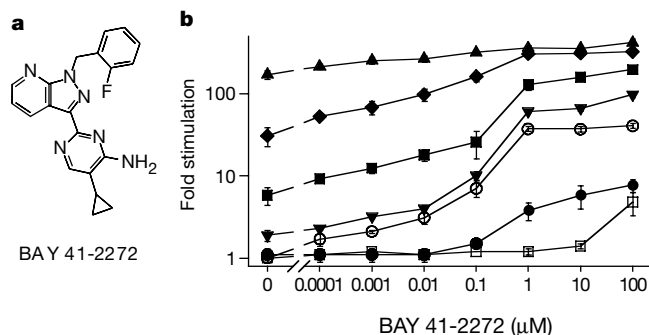


Figure 1 *In vitro* effects of BAY 41-2272. **a**, Structure of BAY 41-2272. **b**, Stimulation of purified sGC by BAY 41-2272 (0.0001–100 μ M) in the absence (open circles) and presence of DEA/NO (0.001 μ M, filled 'down' triangles; 0.01 μ M, filled squares; 0.1 μ M, filled diamonds; 1 μ M, filled 'up' triangles) and ODD (10 μ M, filled circles). Stimulation of haem-free sGC by BAY 41-2272 (open squares). The specific activity of sGC is expressed as fold stimulation versus basal activity (basal activity in the presence of Mg^{2+} 146 nmol mg^{-1} min^{-1}). Data represent means $\pm$ s.e.m. from three independent experiments performed in duplicate.

and BAY 41-2272 (100 μ M) examined, the specific activity of sGC was 60 μ mol $\text{mg}^{-1} \text{min}^{-1}$, 416-fold above the baseline. Although BAY 41-2272 alone is not as strong a stimulator of sGC as is NO, concentrations as low as 10–100 nM stimulate sGC to a level that would be expected to cause biologically important increases in cGMP. 1H-(1,2,4)-Oxadiazolo-(4,3a)-6-bromoquinoxalin-1-one (ODQ), a potent and selective inhibitor of sGC^{6–8}, completely inhibited the effect of BAY 41-2272 (Fig. 1b).

NO stimulates sGC through the formation of a nitrosyl-haem complex, indicating that the mechanism is haem dependent³. Thus, removal of the haem group by the non-ionic detergent Tween-20 (0.5 %) led to an NO-insensitive form of sGC without destructing basal enzyme activity⁹. Even BAY 41-2272 did not activate the haem-free enzyme (Fig. 1b). These results show that BAY 41-2272 activates sGC by an NO-independent, but haem-dependent mechanism.

To determine whether BAY 41-2272 interacts directly with the prosthetic haem group, we recorded the ultraviolet–visual spectra of the purified sGC under unstimulated and stimulated conditions¹⁰. NO caused the characteristic shift of the Soret peak to lower wavelength, and the addition of BAY 41-2272 resulted in no change of the Soret band of either the non-stimulated (431 nm) or NO-stimulated (398 nm) enzyme. Unlike NO, therefore, BAY 41-2272 probably does not bind to the haem moiety of sGC.

To determine the binding site of this structural class of compounds, we carried out a photoaffinity labelling study. For this purpose, we synthesized BAY 51-9491, a derivative of BAY 41-2272 with a tritium label and a photolabile azido group (³H-PAL) coupled to the core structure through a benzoylic spacer (Fig. 2a). Similar to BAY 41-2272, BAY 51-9491 stimulated sGC and showed, in combination with NO, a potentiation of sGC activity. sGC stimulation by BAY 51-9491 could be completely blocked by ODQ. Because BAY 51-9491 showed no effects on haem-depleted enzyme and stimulated sGC with similar characteristics to those of BAY 41-2272, we conclude that BAY 51-9491 is also a direct and NO-independent stimulator of sGC.

sGC and ³H-PAL were irradiated in the presence and absence of unlabelled PAL, BAY 41-2272, YC-1 and ODQ. ³H-PAL bound almost exclusively to the α_1 -subunit of the enzyme (Fig. 2b). Unlabelled PAL and BAY 41-2272 and YC-1 diminished the covalent binding of ³H-PAL to the α_1 -subunit of sGC, thereby suggesting that they have a common binding site. Although ODQ also inhibited the binding of ³H-PAL to the α_1 -subunit, weak labelling of the β_1 -subunit resulted. As ODQ changes the haem environment of the sGC by oxidation^{7,8}, we propose that the binding site for direct and NO-independent sGC stimulators is destroyed and therefore ³H-PAL binds unspecifically to both subunits of the sGC. We could exclude an unspecific interaction between sGC and ³H-PAL, because ³H-PAL processed with sGC without irradiation showed no insertion of radioactivity (Fig. 2b).

To specify the ³H-PAL-binding site in the α_1 -subunit, we labelled sGC with ³H-PAL and fragmented the labelled protein by CNBr cleavage. The resulting protein fragments were separated by electrophoresis, transferred to a polyvinylidene difluoride (PVDF) membrane, and exposed to imaging plates (Fig. 2c, d). Sequencing of the identified highly labelled protein fragments showed that ³H-PAL binds to CNBr fragment VIII (amino acids 236–290; see Supplementary Information). We determined the radioactivity contained in the single phenylthiohydantoin (PTH) cycles from the sequencing of the CNBr VIII fragment, and found that there was a distinct increase of radioactivity in cycles 3 and 8 (>10-fold over basal measured activity), corresponding to cysteines 238 and 243. We therefore conclude that ³H-PAL is bound to sGC through Cys 238 and Cys 243 in the α_1 -subunit of sGC (Fig. 2e). These cysteines are conserved between the amino-acid sequence of rat and human sGC^{11,12}. The relevance of the identified sequence as a regulatory unit remains to be confirmed by mutational analysis and co-crystallization studies.

Through its interaction with this binding site in sGC, BAY 41-2272 possesses an unique pharmacological *in vitro* and *in vivo* profile. In phenylephrine-precontracted ($3 \times 10^{-8} \text{ g ml}^{-1}$) rabbit aortic rings^{5,13}, BAY 41-2272 elicited a concentration-dependent relaxation with a half-maximal inhibitory concentration (IC_{50}) of $304 \pm 63 \text{ nM}$ ($n = 6$), whereas glycerol trinitrate and YC-1 used as controls had an IC_{50} of $1,300 \pm 385 \text{ nM}$ ($n = 6$) and $10,035 \pm 632 \text{ nM}$ ($n = 26$).

Moreover, sGC stimulation in platelets correlates with inhibition of aggregation, platelet cGMP increase, phosphorylation of vasodilator-stimulated phosphoprotein, and prolongation of bleeding time *in vivo*¹⁴. The aggregation of washed human platelets induced by collagen is not influenced by glycerol trinitrate, whereas the NO donors 3-morpholinodisynonimine (SIN-1) and sodium nitropruside (SNP) blocked aggregation with an IC_{50} of $870 \pm 55 \text{ nM}$ ($n = 5$) and $1,141 \pm 34 \text{ nM}$ ($n = 8$), respectively. BAY 41-2272 is more than one order of magnitude more potent than SNP and SIN-1 (IC_{50} $36 \pm 3 \text{ nM}$; $n = 9$). After oral administration, BAY 41-2272 (0.3 mg per

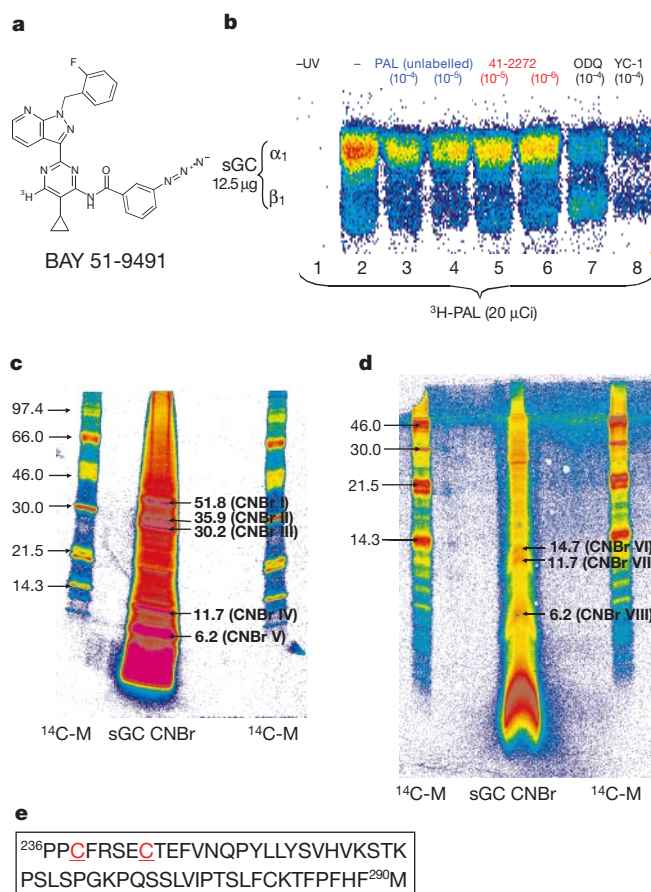


Figure 2 The NO-independent regulatory site of sGC. **a**, Structure of the ³H photoaffinity label (³H-PAL). **b**, Autoradiogram of photoaffinity-labelled sGC after separation by SDS-PAGE. Lane 1, 12.5 μ g sGC with ³H-PAL without irradiation; lane 2, 12.5 μ g sGC with ³H-PAL after irradiation; lane 3, as lane 2 plus 100 μ M unlabelled PAL; lane 4, as lane 2 plus 10 μ M unlabelled PAL; lane 5, as lane 2 plus 10 μ M BAY 41-2272; lane 6, as lane 2 plus 1 μ M BAY 41-2272; lane 7, as lane 2 plus 100 μ M ODQ; lane 8, as lane 2 plus 100 μ M YC-1. **c**, Autoradiogram of the western blot of photoaffinity-labelled sGC fragments after CNBr cleavage and SDS-PAGE (10–20%). Shown are ¹⁴C-labelled rainbow molecular weight markers (¹⁴C-M) and photoaffinity-labelled sGC fragments (BrCN I–V) with relative molecular masses (in thousands), as calculated after sequencing. **d**, Autoradiogram of the western blot of photoaffinity-labelled sGC fragments after CNBr cleavage and separation by Tris-Tricine gel. Shown are ³H-PAL-labelled sGC fragments (BrCN VI–VIII) as in **c**. **e**, Sequence alignment of the BrCN VIII fragment of the α_1 -subunit of sGC. The binding sites of ³H-PAL are shown in red and underlined.

kg (body weight), 172 ± 3 s; 1.0 mg per kg, 186 ± 5 s; 3.0 mg per kg, 191 ± 3 s; controls, 95 ± 3 s) and acetylsalicylic acid used as positive control (30 mg/kg, 212 ± 8 s) ($n = 10$ per group) significantly prolonged tail-bleeding time in rats. These results show that BAY 41-2272 has potential as a representative of a new class of anti-aggregatory drugs.

We and others have shown that YC-1 activates purified sGC and sensitizes the enzyme for NO *in vitro*, in human platelets, on a sGC-overexpressing cell line and in smooth muscle cells^{15–20}. Compared with BAY 41-2272, about 100-fold higher concentrations of YC-1 would be needed to give similar stimulation at sGC^{9,10,15}. The structural similarity between BAY 41-2272 and YC-1 might suggest that both compounds act by the same mechanism. In contrast to YC-1 (ref. 21), however, BAY 41-2272, up to 10^{-5} M, is devoid of any PDE-5 inhibitory activity. Therefore, the therapeutic potential of pure sGC stimulation through this new mechanism with this specific and potent compound might be verified *in vivo* for the first time.

Oral administration of BAY 41-2272 (1–10 mg per kg (body weight)) to hypertensive rats equipped with a radiotelemetric device for the continuous recording of haemodynamic parameters induced a dose-dependent and long-lasting decrease in mean

arterial blood pressure (Fig. 3a). The cardiovascular consequences of stimulation of sGC were evaluated by determining the compound's long-term effects in a high-renin, low-NO rat model of hypertension. Transgenic renin rats with an additional renin gene (TG(mRen2)27) represent a very sensitive model for the cardiovascular effects of NO-synthase inhibition²². Systolic blood pressure measured by the tail cuff method increased in 18-week-old TG(mRen2)27 rats receiving the NO synthase inhibitor L-NAME (N ω -nitro-L-arginine methylester) from 211 ± 5 to 281 ± 2 mm Hg, whereas in animals treated additionally with BAY 41-2272, the increase in systolic pressure could be completely prevented without any signs of tolerance (Fig. 3b). The overall beneficial effects of BAY 41-2272 manifested as both antiplatelet effect and vasodilation were reflected in a significant reduction in mortality. During the study, 13 out of 19 control animals died, whereas in the treatment group only 2 out of 20 animals died ($P < 0.05$, χ^2 -test).

It has been proposed that YC-1 binds to an allosteric site on sGC, and thereby increases the maximal catalytic rate and sensitizes the enzyme towards its gaseous activators NO and CO^{11,12,23}. Studies of the kinetics and equilibria of sGC in the presence of YC-1 has led to a mechanistic model, which attributes a crucial role to the proximal bond that connects the haem iron to histidine 105 of the β -subunit of sGC^{23,24}, but also requires protein control of the distal environment²⁵. In addition, it has been postulated²⁰ that there is an intramolecular sixth ligand of the haem, which oscillates on and off the sixth coordinate, thereby conferring some sort of ligand specificity (that is, NO and CO). The binding of direct sGC stimulators to this site might block this intramolecular binding site and so strengthen the binding of NO and CO to the enzyme. This model is concomitant with the sensitizing of the enzyme towards NO by this new class of stimulators.

In summary, the effects of BAY 41-2272 on sGC and the photo-affinity labelling studies suggest the existence of a new NO-independent regulatory site on sGC in the Cys 238 and Cys 243 region of the α_1 -subunit that modulates the catalytic rate and the responsiveness towards the haem ligand. Our data offer both an approach to understanding the regulation of sGC and a potent new stimulator of sGC, BAY 41-2272, which induces vasodilation without developing nitrate tolerance, antiplatelet activity, and finally reduces mortality. □

Methods

Compounds

BAY 41-2272 (5-Cyclopropyl-2-[1-(2-fluoro-benzyl)-1H-pyrazolo[3,4-b]pyridin-3-yl]-pyrimidin-4-ylamine) and YC-1 (3-(5'-Hydroxymethyl-2'-furyl)-1-benzyl-indazole) were synthesized as described^{5,10}. For 3 H-PAL (BAY 51-9491), 3-azido-N-[5-cyclopropyl-2-[1-(2-fluorobenzyl)-1H-pyrazolo[3,4-b]pyridin-3-yl]-4-[6- 3 H]-pyridinyl]benzamide, the tritium label was introduced on BAY 41-2272 by catalytic dehalogenation of a brominated precursor². Coupling with 3-azidobenzoic acid led to 3 H-PAL with a specific activity of 7.9 Ci mmol^{-1} ($292 \text{ GBq mmol}^{-1}$).

sGC assay

We purified sGC by using a baculovirus/Sf9 expression system and measured enzyme activity in the presence of Mg^{2+} as described¹⁰.

Platelet aggregation

We prepared washed human platelets as described¹⁴, and measured platelet aggregation by the turbidometric method²⁶. The platelet suspension was pre-incubated with the test compound at 37°C for 10 min, and collagen ($0.1\text{--}2 \mu\text{g ml}^{-1}$) was added to induce platelet aggregation.

Blood pressure in conscious, hypertensive rats

Female spontaneously hypertensive rats (200–250 g) were equipped with implantable radiotelemetry, and a data acquisition system (Data Sciences Inc., St. Paul, MN, USA), comprising a chronically implantable transducer/transmitter unit equipped with a fluid-filled catheter, was used. The transmitter was implanted into the peritoneal cavity and the sensing catheter was inserted into the descending aorta. The substance was administered in 10/20/70 (v/v/v) Transcutol/Cremophor EL/water by oral gavage in a volume of 1 mg per kg (body weight).

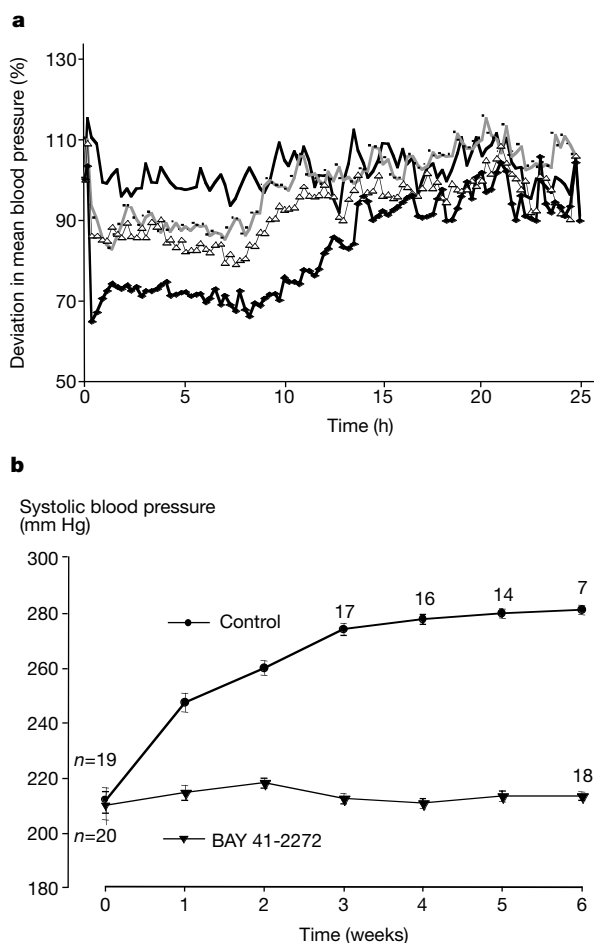


Figure 3 *In vivo* effects of BAY 41-2272. **a**, Effect of orally administered BAY 41-2272 (0 mg per kg (body weight), solid line; 1 mg per kg, filled circles; 3 mg per kg, open triangles; 10 mg per kg, filled diamonds) on mean arterial blood pressure of spontaneously hypertensive rats. Values depicted represent changes in mean arterial blood pressure ($n = 6$ per group). The mean blood pressure at baseline was 149 ± 10 , 148 ± 9 , 148 ± 12 and 151 ± 6 mm Hg, respectively. **b**, Development of systolic blood pressure in TG(mRen2)27 rats given 500 mg l^{-1} L-NAME in their drinking water that were either treated orally with 10 mg per kg (body weight) of BAY 41-2272 or left untreated. Numbers over or under the curve indicate the number of surviving animals.

Labelling of purified sGC

sGC (12.5 µg; purity about 91%) was dissolved in PAL buffer (50 mM TEA-HCl, 0.1 mM EGTA, 1 mM cGMP, 3 mM MgCl₂, 200 µM GTP, pH 7.4) and incubated with 20 µCi ³H-PAL (25 µM) in the presence or absence of different sGC modulators (5 min, 37 °C). Samples were irradiated at 254 nm (N8K ultraviolet 254-nm hand lamp, CAMAG, Berlin, Germany; distance 30 mm, 20 °C, 15 min) and the reaction was stopped by adding SDS-containing stop solution (312.5 mM Tris-HCl, 10% (w/v) SDS, 50% (v/v) glycerol, 250 mM dithiothreitol (DTT), 0.025% (w/v) bromophenol blue, pH 6.8) and heating (5 min at 80 °C). For control 12.5 µg sGC was incubated with 20 µCi ³H-PAL without irradiation. Separation was performed on a 7.5% SDS-PAGE (Protein II cell; Bio-Rad, München, Germany) using a modified Laemmli method²⁷. Proteins were fixed for 20 min in 30/10/60 methanol/acetic acid/MilliQ water, dried and exposed to BAS-TR 20/25 imaging plates (Fuji Photo Film; Tokyo, Japan) for 15 days²⁸. After exposure, the Imaging Plates were scanned by means of a laser scanner (BAS 5000 Scanner; RayTest, Straubenhardt, Germany). Evaluation was performed by visual classification of radiographic intensities, displayed by use of pseudo-colours, starting with blue for lowest detectable concentrations up to red for highest concentrations.

CNBr fragmentation of labelled sGC

sGC (200 µg) was labelled with 1,600 µCi ³H-meta-PAL, and the reaction was terminated by adding TCA to a final concentration of 10%. The protein was precipitated, washed with ice-cold ethanol/ether (1/1), dissolved in formic acid (70% (v/v)) and reacted with a few crystals of CNBr in the dark (24 h, 20 °C) under oxygen-free nitrogen. After evaporation in a Speed Vac (Bachofar, Reutlingen, Germany) for three times after the addition of 1 ml MilliQ water, the pellet was dissolved in sample buffer (62.5 mM Tris-HCl, 2% (w/v) SDS, 10% (v/v) glycerol, 50 mM DDT, 0.0025% (w/v) bromophenol blue, pH 6.8) heated (5 min, 80 °C), and the protein fragments were separated on a 10–20% gradient SDS-PAGE or a 16.5% Tris-Tricine gel. The digested protein fragments were transferred to a PVDF membrane (Trans-Blot Electrophoretic Transfer Cell; Bio-Rad) and stained with Coomassie blue. Dried membranes were exposed to BAS-TR 20/25 Imaging Plates for 6 h without (10–20% gradient gel blot) or for 48 h with a screen of metal foil for contamination monitors (0.9 mg cm⁻² Steiner; Erndtebrück-Schameder, Germany) (16.5% Tris-Tricine gel blot).

Amino-terminal sequence analysis

Amino-terminal sequence analyses were performed using the gas-liquid-solid-phase protein sequencer Procise with an RP-18-PTH-column from Applied Biosystems (Forster City, CA, USA).

Sequence analysis

PVDF-membrane pieces were washed two times with 100 µl of 50% MeOH before sequencing. The cyanogen bromide fragment VIII was sequenced over 60 cycles using the sequencer program for blot sequencing. The single PTH-amino-acids were collected, lyophilized and radioactivity was counted after combustion (Oxidizer model 307, Oximate 80; Packard, Illinois, USA) in a liquid scintillation counter (LS-6500; Beckman, Munich, Germany).

Statistics

Results are given as the mean ± s.e.m. Differences were assessed by one-way analysis of variance (ANOVA) followed by Bonferroni test for comparison of means.

Received 18 April; accepted 11 December 2000.

- Moncada, S., Palmer, R. M. & Higgs, E. A. Nitric oxide: physiology, pathophysiology, and pharmacology. *Pharmacol. Rev.* **43**, 109–142 (1991).
- Furchgott, R. F. Endothelium-derived relaxing factor: Discovery, early studies, and identification as nitric oxide (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **38**, 1870–1880 (1999).
- Murad, F. Discovery of some of the biological effects of nitric oxide and its role in cell signaling (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **38**, 1857–1868 (1999).
- Ignarro, L. J. Nitric oxide: A unique endogenous signaling molecule in vascular biology (Nobel lecture). *Angew. Chem. Int. Ed. Engl.* **38**, 1882–1892 (1999).
- Straub, A. *et al.* Substituted pyrazole derivatives. WO 2,000,006,568 (2000). *Chem. Abstr.* **132**, 122629 (2000).
- Garthwaite, J. *et al.* Potent and selective inhibition of nitric oxide-sensitive guanylyl cyclase by 1H[1,2,4]oxadiazolo[4,3-a]quinoxalin-1-one. *Mol. Pharmacol.* **48**, 184–188 (1995).
- Schrammel, A., Behrends, S., Schmidt, K., Koesling, D. & Mayer, B. Characterization of 1H[1,2,4]oxadiazolo[4,3-a]quinoxalin-1-one as a heme-site inhibitor of nitric oxide-sensitive guanylyl cyclase. *Mol. Pharmacol.* **50**, 1–5 (1996).
- Brunner, F., Schmidt, K., Nielsen, E. B. & Mayer, B. Novel guanylyl cyclase inhibitor potently inhibits cyclic GMP accumulation in endothelial cells and relaxation of bovine pulmonary artery. *J. Pharmacol. Exp. Ther.* **277**, 48–53 (1996).
- Friebe, A., Schultz G. & Koesling, D. Sensitizing soluble guanylate cyclase to become a highly CO-sensitive enzyme. *EMBO J.* **15**, 6863–6868 (1996).
- Hoenicka, M. *et al.* Purified soluble guanylyl cyclase expressed in a baculovirus/Sf9 system: stimulation by YC-1, NO and CO. *J. Mol. Med.* **77**, 14–23 (1999).
- Nakane, M. *et al.* Molecular cloning and expression of cDNAs coding for soluble guanylate cyclase from rat lung. *J. Biol. Chem.* **265**, 16841–16845 (1990).
- Zabel, U., Weeger, M., La, M. & Schmidt, H. H. W. Human soluble guanylate cyclase: functional expression and revised isoenzyme family. *Biochem. J.* **333**, 51–57 (1998).
- Stasch, J. P., Kazda, S. & Neuser, D. Different effects of ANP and nitroprusside on cyclic GMP extrusion of isolated aorta. *Eur. J. Pharmacol.* **174**, 279–282 (1989).

- Becker, E. M. *et al.* The vasodilator-stimulated phosphoprotein (VASP): target of YC-1 and nitric oxide effects in human and rat platelets. *J. Cardiovasc. Pharmacol.* **35**, 390–397 (2000).
- Mulsch, A. *et al.* Effect of YC-1, an NO-independent, superoxide-sensitive stimulator of soluble guanylyl cyclase, on smooth muscle responsiveness to nitrovasodilators. *Br. J. Pharmacol.* **120**, 681–689 (1997).
- Ko, F. N., Wu, C. C., Kuo, S. C., Lee, F. Y. & Teng, C. M. YC-1, a novel activator of platelet guanylate cyclase. *Blood* **84**, 4226–4233 (1994).
- Wu, C. C., Ko, F. N., Kuo, S. C., Lee, F. Y. & Teng, C. M. YC-1 inhibited human platelet aggregation through NO-independent activation of soluble guanylate cyclase. *Br. J. Pharmacol.* **116**, 1973–1978 (1995).
- Becker, E. M. *et al.* Generation and characterization of a stable soluble guanylate cyclase-over-expressing CHO cell line. *Nitric Oxide* **3**, 55–66 (1999).
- Friebe, A. & Koesling, D. Mechanism of YC-1 induced activation of soluble guanylyl cyclase. *Mol. Pharmacol.* **53**, 123–127 (1998).
- Hobbs, A. J. Soluble guanylate cyclase: the forgotten sibling. *Trends Pharmacol. Sci.* **18**, 484–491 (1997).
- Galle, J. *et al.* Effects of the soluble guanylyl cyclase activator, YC-1, on vascular tone, cyclic GMP levels and phosphodiesterase activity. *Br. J. Pharmacol.* **127**, 195–203 (1999).
- Hirth-Dietrich, C., Stasch, J. P. & Ganten, D. Transgenic rats with an additional renin gene (TGR(mRen2)27) as a sensitive model for the renal effects of NO synthase inhibition. *Naunyn-Schmiedeberg's Arch. Pharmacol.* **350**, R6 (1994).
- Zhao, Y., Schelvis, J. P. M., Babcock, G. T. & Marletta, M. A. Identification of histidine 105 in the β1 subunit of soluble guanylate cyclase as the heme proximal ligand. *Biochemistry* **37**, 4502–4509 (1998).
- Zhao, Y. & Marletta, M. A. Localization of the heme binding region in soluble guanylate cyclase. *Biochemistry* **36**, 15959–15964 (1997).
- Kharitonov, V. G., Sharma, V. S., Magde, D. & Koesling, D. Kinetics and equilibria of soluble guanylate cyclase ligation by CO: effect of YC-1. *Biochemistry* **38**, 10699–10706 (1999).
- Born, G. V. R. & Cross, M. J. The aggregation of blood platelets. *J. Physiol. (Lond.)* **168**, 178–185 (1963).
- Laemmli, U. K. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685 (1970).
- Ahr, H. H. & Steinke, W. Vasodilation and first experience with quantitative studies in whole-body autoradiography during drug development. *Xenobiotic Metab. Dispos.* **9**, 371–378 (1994).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank C. Robyr-Fürstner for the synthesis of YC-1; E. Bischoff for performing the PDE assays; and D. Wood for critical comments on the manuscript.

Correspondence and requests for materials should be addressed to J.-P.S. (e-mail: johannes-peter.stasch.js@bayer-ag.de).

Polo-like kinase 1 phosphorylates cyclin B1 and targets it to the nucleus during prophase

Fumiko Toyoshima-Morimoto^{*†}, Eri Taniguchi[†], Nobuko Shinya[‡], Akihiro Iwamatsu[‡] & Eisuke Nishida^{*†}

^{*} Department of Biophysics, Graduate School of Science, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan

[†] Department of Cell and Developmental Biology, Graduate School of Biostudies, Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan

[‡] Central Laboratories for Key Technology, Kirin Brewery Company Limited, Kanazawa-ku, Yokohama-shi, Kanagawa 236-0004, Japan

In vertebrate cells, the nuclear entry of Cdc2–cyclin B1 (MPF^{1–3}) during prophase^{4–6} is thought to be essential for the induction and coordination of M-phase events^{7–12}. Phosphorylation of cyclin B1 is central to its nuclear translocation^{8,13,14}, but the kinases that are responsible remain unknown. Here we have purified a protein kinase from *Xenopus* M-phase extracts that phosphorylates a crucial serine residue (S147) in the middle of the nuclear export signal sequence^{10,13,15} of cyclin B1. We have identified this kinase as Plx1 (ref. 16), a *Xenopus* homologue of Polo-like kinase (Plk)-1 (refs 17, 18). During cell-cycle progression in HeLa cells, a change in the kinase activity of endogenous Plk1 toward S147 and/or S133

correlates with a kinase activity in the cell extracts. An anti-Plk1 antibody depletes the M-phase extracts of the kinase activity toward S147 and/or S133. An anti-phospho-S147 antibody reacts specifically with cyclin B1 only during G2/M phase. A mutant cyclin B1 in which S133 and S147 are replaced by alanines remains in the cytoplasm, whereas wild-type cyclin B1 accumulates in the nucleus during prophase. Co-expression of constitutively active Plk1 stimulates nuclear entry of cyclin B1. Our results indicate that Plk1 may be involved in targeting MPF to the nucleus during prophase.

When vertebrate cells enter prophase, cyclin B1 translocates from the cytoplasm to the nucleus^{4,5,19}. One of the putative roles of nuclear MPF as the nuclear lamin kinase is crucial for nuclear envelope breakdown at the end of prophase²⁰. Four serine residues (S94, S96, S101 and S113) of *Xenopus* cyclin B1 have been identified as *in vivo* phosphorylation sites during the G2/M transition^{21,22}, and their phosphorylation is thought to be responsible for both nuclear entry of cyclin B1 and germinal vesicle breakdown^{8,13,14,22}.

When an haemagglutinin A (HA)-tagged mutant of human cyclin B1, in which four serine residues (S126, S128, S133, S147) were replaced by alanine (S126,128,133,147A-cyclin B1), was co-expressed with Myc-tagged wild-type cyclin B1 in HeLa cells, the mutant cyclin B1 remained in the cytoplasm, whereas the wild type accumulated in the nucleus in prophase (Fig. 1a, middle). Myc-tagged wild-type cyclin B1 was indistinguishable from HA-tagged wild-type cyclin B1 in nuclear accumulation during prophase (Fig. 1a, top). This result is consistent with a report that a non-

phosphorylatable form of cyclin B1 can enter the nucleus only after nuclear envelope breakdown¹⁴. Not only S126,128,133,147A-cyclin B1, but also S133,147A-cyclin B1 remained in the cytoplasm during prophase, whereas wild-type cyclin B1 accumulated in the nucleus (Fig. 1a, bottom). S126,128A-cyclin B1 was not exclusively cytoplasmic, and localized, but not accumulated, in the nucleus during prophase (data not shown).

We constructed a series of mutant forms of cyclin B1, in which these four serines were replaced by glutamate to mimic phosphorylation. When each of the serines, or the first two or the last two, were replaced by glutamate, cyclin B1 remained in the cytoplasm (Fig. 1b, S126E, S128E, S133E, S147E, S126,128E and S133,147E). When all four serines were replaced by glutamate, cyclin B1 localized in the nucleus (Fig. 1b, S126,128,133,147E). Together, these results suggest that phosphorylation of S133 and S147 is necessary for the nuclear translocation of cyclin B1 during prophase, and that phosphorylation of S126 and S128 may stimulate the nuclear translocation.

MPF itself can phosphorylate the first two serines^{14,21,23}, although it has not been proved that MPF regulates nuclear localization of cyclin B1. As our above results showed the importance of identifying a protein kinase(s) that phosphorylates the last two serines, we prepared extracts from synchronized HeLa cells and assayed them for kinase activities toward the various forms of cyclin B1. Kinase activity toward S126,128A-cyclin B1 and toward wild-type cyclin B1 increased roughly in parallel with the activity of MPF—assayed by histone H1 kinase activity—and decreased slightly later than that of MPF (Fig. 2a). Kinase activities toward S126,128,133,147A-cyclin B1 were not observed (Fig. 2a). Thus, a kinase activity that phosphorylates S133 and/or S147 of cyclin B1 becomes activated during G2/M transition in HeLa cells.

U0126, a specific inhibitor of mitogen-activated protein (MAP) kinase kinase MEK at a concentration enough to inhibit epidermal growth factor (EGF)-induced activation of MAP kinase ERK in HeLa cells, did not block the kinase activity of the extracts toward S126,128A-cyclin B1 (Fig. 2b). This indicates that MAP kinase may not be involved in phosphorylating S133 and/or S147 of cyclin B1 during G2 and M phases in HeLa cells. A kinase activity toward S126,128A-cyclin B1 was also detected in the process of *Xenopus* oocyte maturation and found to increase after progesterone treatment (Fig. 2c). Thus, a kinase activity that phosphorylates S133 and/or S147 of cyclin B1 is also activated during G2/M transition in *Xenopus* oocyte maturation.

We purified from *Xenopus* M-phase extracts a kinase activity

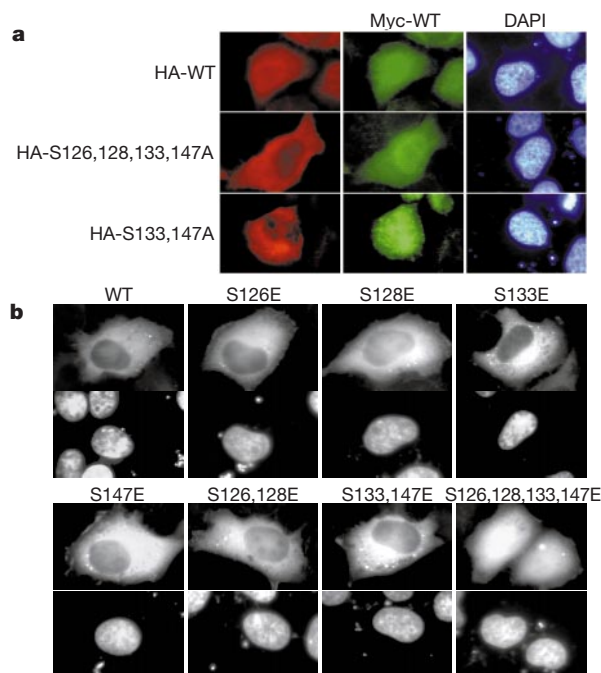


Figure 1 Phosphorylation-dependent nuclear localization of cyclin B1. **a**, Myc-tagged or HA-tagged wild-type (WT) cyclin B1, HA-S126,128,133,147A-cyclin B1, or HA-S133,147A-cyclin B1 were co-expressed in HeLa cells. An image of a prophase cell is shown. **b**, Top, subcellular localization of the indicated forms of HA-cyclin B1 (top). Bottom, DAPI staining. Quantification of staining in the cells is as follows, where N > C indicates staining intensity in the nucleus (N) is stronger than that in the cytoplasm (C); N = C indicates N is equal to C; and N < C indicates C is stronger than N. WT (total 202 cells): N > C, 0; N = C, 1; N < C, 201. S126E (total 158 cells): N > C, 0; N = C, 1; N < C, 157. S128E (total 185 cells): N > C, 0; N = C, 0; N < C, 185. S133E (total 164 cells): N > C, 0; N = C, 0; N < C, 164. S147E (total 190 cells): N > C, 0; N = C, 2; N < C, 188. S126,128E (total 202 cells): N > C, 0; N = C, 21; N < C, 181. S133,147E (total 203 cells): N > C, 1; N = C, 17; N < C, 185. S126,128,133,147E (total 229 cells): N > C, 204; N = C, 20; N < C, 5.

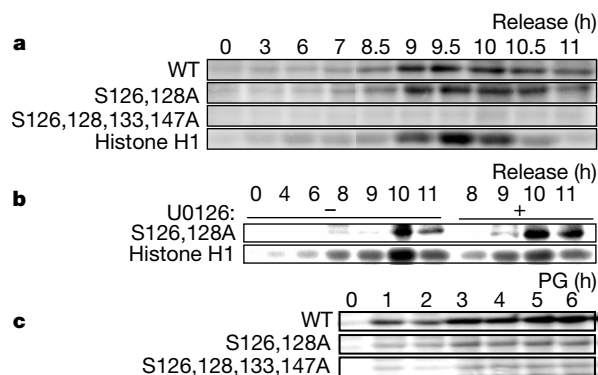


Figure 2 A kinase activity that phosphorylates S133 and/or S147 of cyclin B1 is activated during the G2/M transition. **a**, **b**, Kinase activities in the extracts from synchronized HeLa cells by a double thymidine block²⁰ toward the indicated substrates. At 6 h after release, U0126 (10 μ M) was added in the assays in **b**. **c**, Stage VI *Xenopus* oocytes were treated with progesterone (PG). Oocyte extracts were prepared at the indicated times after treatment with progesterone and the kinase activities in the extracts were measured as in **a**.

toward S126,128A-cyclin B1, which eluted in the flow-through fraction after chromatography on Q-Sepharose (Fig. 3a). We purified the flow-through fraction by sequential chromatography on glutathione *S*-transferase (GST)–cyclinB1–Sepharose, heparin–Sepharose and oligo(dT)–cellulose; and in all steps the kinase activity toward S126,128A-cyclin B1 eluted as a single peak (Fig. 3b

and Table 1). Elution of a polypeptide of relative molecular mass 65,000 (M_r 65K), the main band detected by silver staining, coincided completely with that of the kinase activity in the final step (Fig. 3b, Oligo(dT)–Cellulose, top and bottom).

The 65K protein was digested with achromobacter protease I. Molecular mass analysis of the resulting Lys-C fragments determined that the 65K protein is Plx1, a *Xenopus* homologue of Plk1. Anti-Plk1 antibody reacted with the 65K protein (Fig. 3b, Oligo(dT)–Cellulose, Plk1 IB), confirming it as Plx1. This result indicates that Plx1 and Plk1 may phosphorylate S133 and/or S147 of cyclin B1 in *Xenopus* oocyte maturation and in the G2/M transition of mammalian cells, respectively.

In agreement with previous reports^{24,25}, both the concentration of Plk1 and the total kinase activity of Plk1 toward casein, a substrate of Plk1, increased during G2/M phase in synchronized HeLa cells (Fig. 4a, rows 5 and 6). The kinase activity of endogenous Plk1, immunoprecipitated with anti-Plk1 antibody, toward S126,128A-cyclin B1 increased in parallel with the total kinase activity of Plk1 toward casein (Fig. 4a, rows 3 and 5). This profile coincided with the kinase activity of total lysates toward S126,128A-cyclin B1 (Fig. 4a, row 1). Immunoprecipitated Plk1 barely phosphorylated S126,128,133,147A-cyclin B1 (Fig. 4a, row 4). Thus, during G2/M phase Plk1 phosphorylates cyclin B1 on S133 and/or S147 in HeLa cells.

The S126,128A-cyclin B1-phosphorylating activity in the extracts from M-phase HeLa cells was decreased to less than 10% of its original value by immunodepletion with increasing amounts of anti-Plk1 antibody, but not with control IgG (Fig. 4b, top). Bacterially produced recombinant His-tagged Plk1 protein restored the kinase activity of the Plk1-immunodepleted extract (Fig. 4b, bottom, lanes 3 and 4). These results indicate that Plk1 is a major kinase that phosphorylates S133 and/or S147 of cyclin B1 during M phase.

Recombinant His–Plk1 phosphorylated S126,128A-cyclin B1 in a dose-dependent manner, and the efficiency of the phosphorylation was almost the same as that of phosphorylating casein (Fig. 5a), indicating that cyclin B1 is a good substrate for Plk1. We next examined which of the serines S133 or S147, or both, was phosphorylated by Plk1. The active fraction of heparin–Sepharose from mature *Xenopus* oocytes (Fig. 3b) could phosphorylate S126,128A-cyclin B1 and S126,128,133A-cyclin B1, but not S126,128,147A-cyclin B1 (Fig. 3c). Recombinant His–Plk1 protein could phosphorylate S126,128A-cyclin B1 and S126,128,133A-cyclin B1, but not S133,147A-cyclin B1, S126,128,147A-cyclin B1, or S126,128,133,147A-cyclin B1 (Fig. 5b, top). Flag-tagged Plk1, purified by immunoprecipitation with anti-Flag antibody from extracts of Flag-tagged Plk1-transfected HeLa cells, also showed essentially the same substrate specificity (Fig. 5b, bottom). Moreover, recombinant His–Plk1 protein phosphorylated a synthetic peptide corresponding to residues 141–154, DLCQAFSDVILAVN, of cyclin B1 (OVA–NES), but not a phosphopeptide (DLCQAF-phospho^{S147}–DVILAVN; OVA–P–NES) (Fig. 5c). Together, these results show that Plk1 directly phosphorylates S147 of cyclin B1.

Our results with serine/alanine mutants and serine/glutamate mutants of cyclin B1 (Fig. 1a, b) suggest that phosphorylation of S126 and S128, and of S133 and S147 cooperates to induce nuclear localization of cyclin B1. To test whether phosphorylation of S147 in

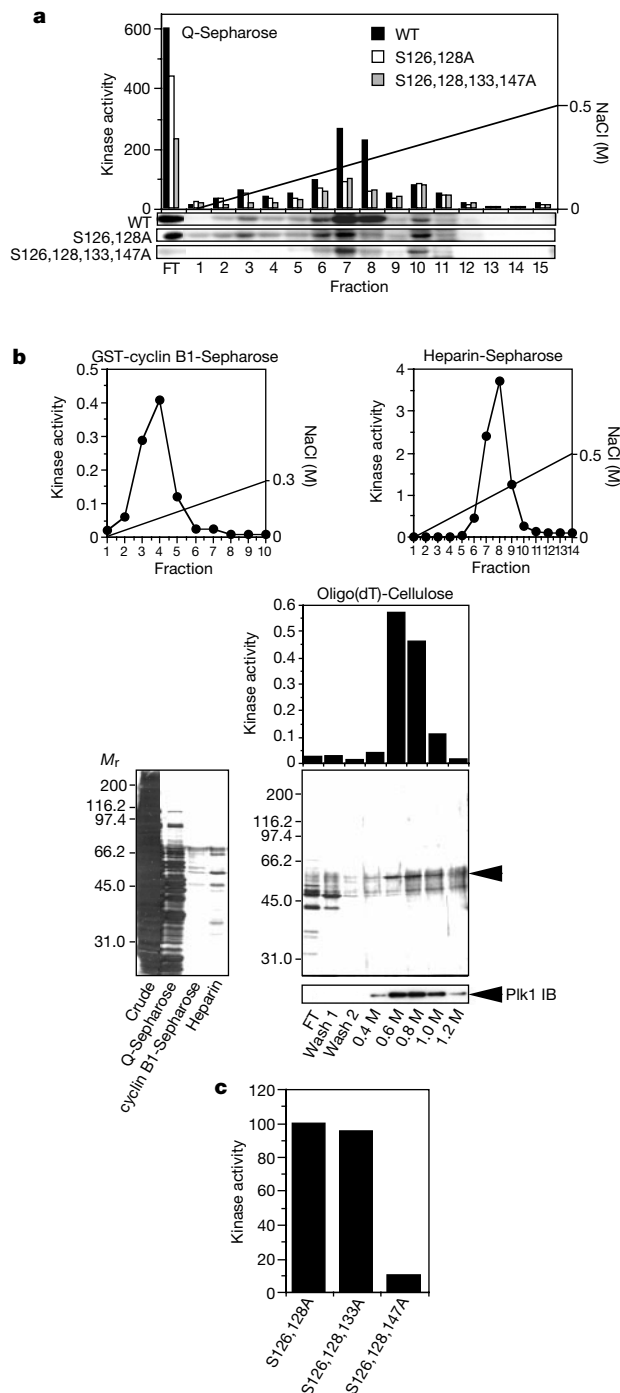


Figure 3 Purification of a kinase activity that phosphorylates S133 and/or S147 of cyclin B1. **a**, Elution profile from Q-Sepharose of the total kinase activities toward indicated substrates (top). Images of the autoradiography (bottom). **b**, Elution profiles of a kinase activity toward S126,128A-cyclin B1 (top and upper middle panels). Data are shown in arbitrary units (1 unit is $0.1 \text{ nmol min}^{-1} \text{ mg}^{-1}$). Silver-staining of each purification step (lower middle panels) and immunoblotting with anti-Plk1 antibody (bottom panel). **c**, Kinase activity in the active fraction (fraction 8) of heparin–Sepharose toward the indicated forms of cyclin B1.

Table 1 Purification of cyclin B1 kinase from *Xenopus* M-phase extracts

	Total protein (mg)	Total activity (units)	Specific activity (units per mg)	Purification (fold)
Crude extract	220			
Q-Sepharose	53	467	8.8	1.00
GST-cyclin B1-Sepharose	0.33	131	400	50.3
Heparin-Sepharose	0.050	130	2,600	292
Oligo(dT)-Cellulose	0.0015	22.8	15,200	1,710

addition to phosphorylation of the first two serines could induce nuclear localization of cyclin B1, we examined the localization of S126,128,147E-cyclin B1 and found that it localized to the nucleus (Fig. 5d). Thus, phosphorylation of the first two serines (S126 and S128) and S147 should be sufficient for nuclear translocation of cyclin B1 in prophase.

To examine whether phosphorylation of S147 of cyclin B1 occurs during prophase *in vivo*, we produced an antibody that reacts specifically with the S147-phosphorylated cyclin B1. The affinity-purified antibody (147P) recognized the synthetic phosphopeptide (OVA-P-NES), but not a non-phosphopeptide (OVA-NES) (Fig. 5e, top). It reacted with the cyclin B1 that had been immunoprecipitated with a monospecific anti-cyclin B1 antibody (Fig. 5e, middle, extract) from M-phase extracts, but it did not react with the cyclin B1 from G1/S phase (Fig. 5e, middle, anti-cyclin B1 IP). When 147P was mixed with M-phase extracts and the mixture was immunoprecipitated with the anti-cyclin B1 antibody, almost all of the 147P antibody precipitated with cyclin B1 (Fig. 5e, lower, IP; anti-cyclin B1, M). In contrast, the mixed 147P remained in the supernatant after immunoprecipitation with control IgG or when mixed with G1/S phase extracts (Fig. 5e, lower, IP; anti-cyclin B1 and IP; IgG). These data indicate that 147P recognizes specifically the S147-phosphorylated cyclin B1.

The 147P antibody immunoprecipitated cyclin B1 from synchronized HeLa cells during G2/M phase (8–10 h after release), but not from the cells in S phase (4 h after release) (Fig. 5f, bottom, IP; anti-147P) despite accumulation of significant amounts of cyclin B1 in these cells (Fig. 5f, extracts). The change in the amounts of cyclin B1 immunoprecipitated with 147P during the cell cycle matched completely the change in the kinase activity of the extracts to phosphorylate S126,128A-cyclin B1, except for a significant kinase activity in the extracts but not the immunoprecipitated cyclin B1

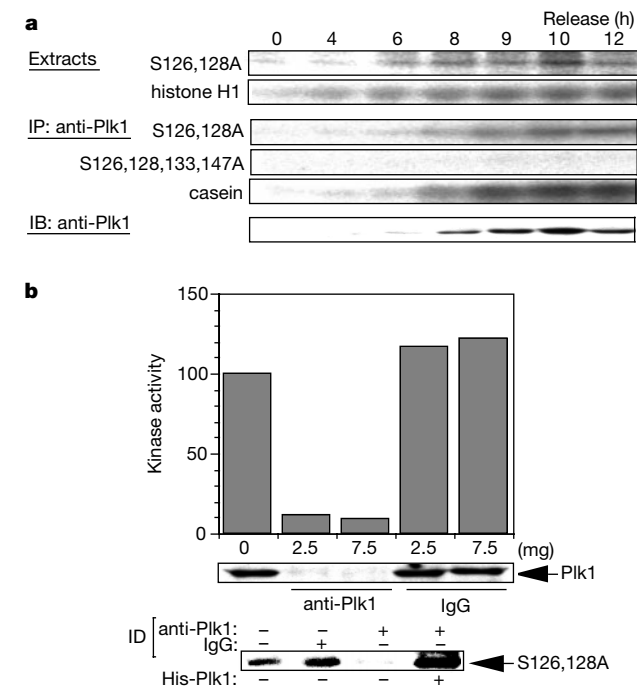


Figure 4 Plk1 phosphorylates cyclin B1 on S133 and/or S147 during G2/M phase in HeLa cells. **a**, Kinase activities in the extracts and that of endogenous Plk1 (IP; anti-Plk1). The extracts were immunoblotted with anti-Plk1 antibody (IB: anti-Plk1). **b**, Immunodepletion (ID) of M-phase extracts with increasing amounts of anti-Plk1 antibody or non-immunized rabbit IgG. Kinase activity toward GST-S126,128A-cyclin B1 remaining in the supernatant (top) and the immunoblotting with anti-Plk1 antibody (middle) are shown. His-tagged Plk1 (0.1 μ g) was added to the Plk1-depleted M-phase extracts (bottom). Kinase activities toward GST-S126,128A-cyclin B1 are shown.

during late M phase, 11 h after release (Fig. 5f, S126, 128A, IP; anti-147P). This may be due to degradation of endogenous cyclin B1 in late M phase (Fig. 5f, bottom).

Immunohistochemistry of HeLa cells with 147P (Fig. 5g) showed that metaphase cells (white arrowheads), but not S-G2 phase (red

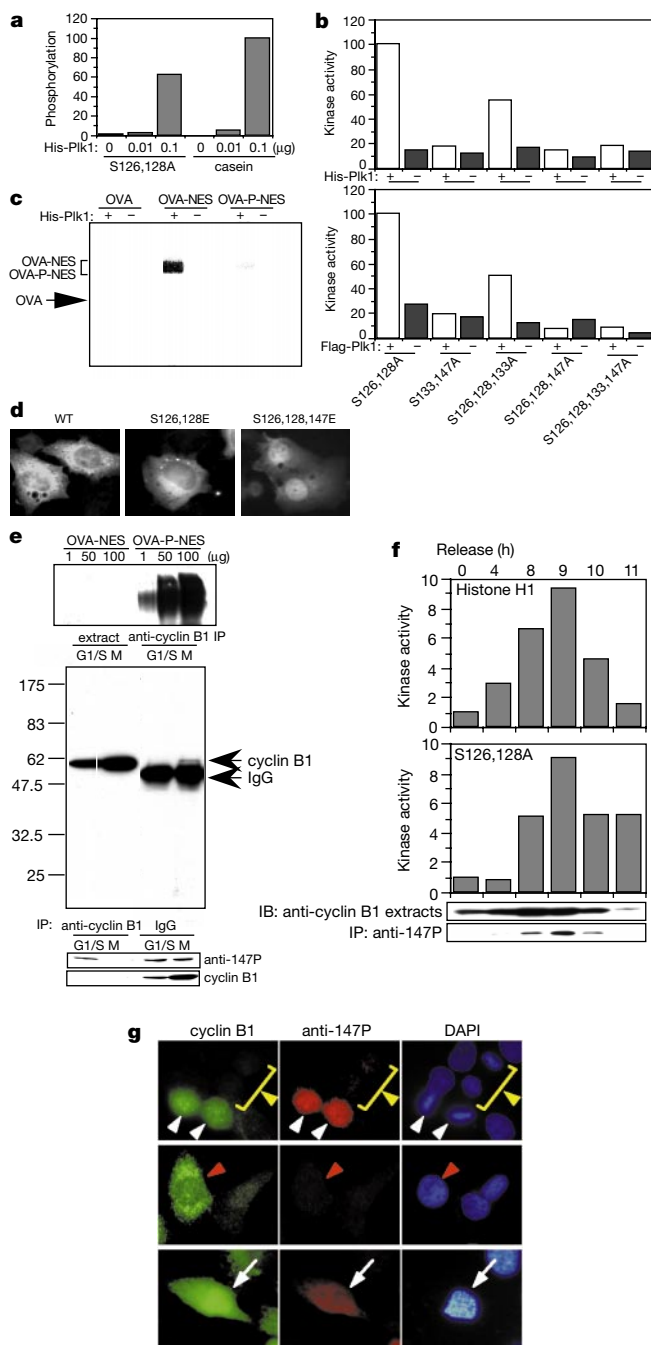


Figure 5 Plk1-mediated phosphorylation of cyclin B1 on S147 *in vitro* and *in vivo*. **a**, Phosphorylation of GST-S126,128A-cyclin B1 and casein (20 pmol of each) by His-Plk1. **b**, Kinase activities of His-Plk1 (0.01 μ g) toward the indicated forms of cyclin B1 (1 μ g). **c**, Kinase activities of His-Plk1 (1 μ g) toward OVA, OVA-NES and OVA-P-NES (0.2 μ g each). Similar results were obtained using 0.1 μ g His-tagged Plk1 and 0.02 μ g of substrates. **d**, Subcellular localization of HA-tagged WT cyclin B1, HA-S126,128E-cyclin B1 and HA-S126,128,147E-cyclin B1. **e**, Specificity of anti-147P (see Methods). **f**, Kinase activities in the extracts toward histone H1 and S126,128-cyclin B1 (top and bottom graphs). Extracts and the immunoprecipitates with anti-147P were immunoblotted with anti-cyclin B1 antibody (bottom panels). **g**, HeLa cells were stained with anti-cyclin B1 antibody (left), anti-147P (middle) or DAPI (right).

arrowhead), anaphase (not shown), telophase (yellow arrowhead) or G1 phase cells (not shown), were stained with this antibody. This antibody stained also prophase cells (white arrow). These results show that phosphorylation of S147 of cyclin B1 occurs during G2/M phase *in vivo*.

Our results indicate that Plk1 phosphorylation of cyclin B1 on S147 during prophase is central to inducing nuclear translocation of cyclin B1. As S126 and S128 of cyclin B1 are reported to be phosphorylated by Cdc2 (refs 14, 21, 23), activation of Plk1 should induce nuclear translocation of S126,128E-cyclin B1. To test this, we co-expressed Plk1 and S126,128E-cyclin B1 in HeLa cells. Experiments were performed in the presence of MG132, a proteasome inhibitor, to avoid degradation of cyclin B1, as Plk1 phosphorylates and activates the anaphase-promoting complex (APC), facilitating degradation of cyclin B1. Treatment with MG132 alone had no effect on the subcellular localization of cyclin B1 (data not shown). When co-expressed with wild-type Plk1, the S126,128E-cyclin B1 remained in the cytoplasm (Fig. 6a, WT). We next co-transfected a constitutively active Plk1 (SDTD-Plk1), constructed by a published method²⁶, and found that the S126,128E-cyclin B1 accumulated strongly in the nucleus (Fig. 6a, SDTD).

To exclude possible involvement of the kinase activity of Cdc2, we carried out experiments in the presence of butyrolactone I, a specific

CDK inhibitor, at a concentration enough to induce cell-cycle arrest at G2 phase but not apoptosis in HeLa cells. S126,128E-cyclin B1 still accumulated strongly in the nucleus when SDTD-Plk1 was co-expressed (Fig. 6a, SDTD+butyrolactone I), indicating that activation of Plk1 can induce nuclear entry of S126,128E-cyclin B1 in the absence of Cdc2 activity. Because expression of SDTD-Plk1 did not induce nuclear translocation of MAP kinase kinase MEK1 (data not shown), whose cytoplasmic localization is maintained by its nuclear export signal (NES), we could exclude the possibility that Plk1 may inhibit the nuclear export system.

To further investigate the role of phosphorylation of cyclin B1 in its nuclear localization, we co-expressed various forms of cyclin B1 with SDTD-Plk1. In most transfected cells (~70–80%), both S126,128,133,147A-cyclin B1 and S133,147A-cyclin B1 remained in the cytoplasm even in the presence of SDTD-Plk1 (Fig. 6b, S126, 128, 133, 147A and S133,147A). This result supports our idea that Plk1 phosphorylates S147 and induces nuclear localization of cyclin B1. In contrast, in more than half of the transfected cells, both wild-type and S126,128A-cyclin B1, localized in the nucleus (Fig. 6b, WT and S126,128A). Thus, the phosphorylation of S126 and/or S128, which is catalysed by Cdc2, may not be absolutely required for the Plk1-induced nuclear localization of cyclin B1. As nuclear localization of wild-type and S126,128A-cyclin B1 was less complete than that of S126,128E-cyclin B1 (Fig. 6b), however, phosphorylation of S126 and/or S128 may also contribute to stimulation of the nuclear localization of cyclin B1.

We have identified Plk1 as a protein kinase that phosphorylates a serine residue (S147) in the NES of cyclin B1 and targets it to the nucleus during prophase in vertebrate cells. The Plk1-mediated phosphorylation seems to be responsible for inactivating the NES of cyclin B1 rather than enhancing nuclear import of cyclin B1, because we observed that OVA-P-NES, when injected into the nucleus, was exported from the nucleus to the cytoplasm more slowly than OVA-NES, and that OVA-P-NES, when injected into the cytoplasm, did not move into the nucleus (data not shown). It is possible, however, that phosphorylation of several serines including S147 may create a nuclear import signal and enhance the nuclear import of cyclin B1 (ref. 14).

During G2/M phase, Cdc2 is activated and Plk1 is synthesized and activated. Thus, the activation of the kinase activity of MPF (Cdc2–cyclin B1 complex) and its nuclear translocation may occur coordinately. Previous work has suggested that Plk1 might activate Cdc2 by phosphorylating and activating Cdc25C (refs 16, 27), and that a positive feedback loop might exist between Plk1 and MPF. We speculate that Plk1 may coordinate the activation of MPF and its nuclear translocation. □

Methods

Protein preparation

GST–cyclin B1 and His-tagged Plk1 were prepared as described¹⁰. His-tagged Plk1 was further purified by Mono-Q with a linear gradient to 1 M NaCl, and the active fractions were used for *in vitro* kinase assays.

Immunoprecipitation, immunoblotting and cell staining

Cells were lysed in buffer 1 (20 mM HEPES pH 7.4, 20 mM 2-glycerophosphate, 2 mM EGTA, 1.5 mM MgCl₂, 150 mM NaCl, 50 mM NaF, 0.5% NP-40, 2 mM dithiothreitol (DTT), 1 mM Na₃VO₄, 1 mM PMSF, 2 μg ml⁻¹ aprotinin, 1 μM okadaic acid) containing 10% glycerol, and centrifuged at 20,000g for 15 min. Endogenous Plk1 or Flag-tagged Plk1 was immunoprecipitated with anti-Plk1 (Zymed) or anti-Flag antibody (Sigma). Immunoblotting and immunofluorescent cell staining were done as described^{10,28}.

Kinase assays

We lysed cells in buffer 2 (20 mM Tris pH 7.5, 60 mM 2-glycerophosphate, 10 mM EGTA, 10 mM MgCl₂, 10 mM NaF, 0.1% NP-40, 1 mM DTT, 1 mM Na₃VO₄, 1 mM PMSF, 2 μg ml⁻¹ aprotinin, 1 μM okadaic acid), and centrifuged them at 20,000g for 15 min. The supernatant was mixed with 1 μg of substrate, 100 μM ATP, 15 mM MgCl₂, 2 μCi of [³²P]-ATP and incubated for 40 min at 25 °C. Histone H1 kinase assay was done as described¹⁰. In the kinase assays for His-tagged Plk1 or immunoprecipitates, 0.01 μg,

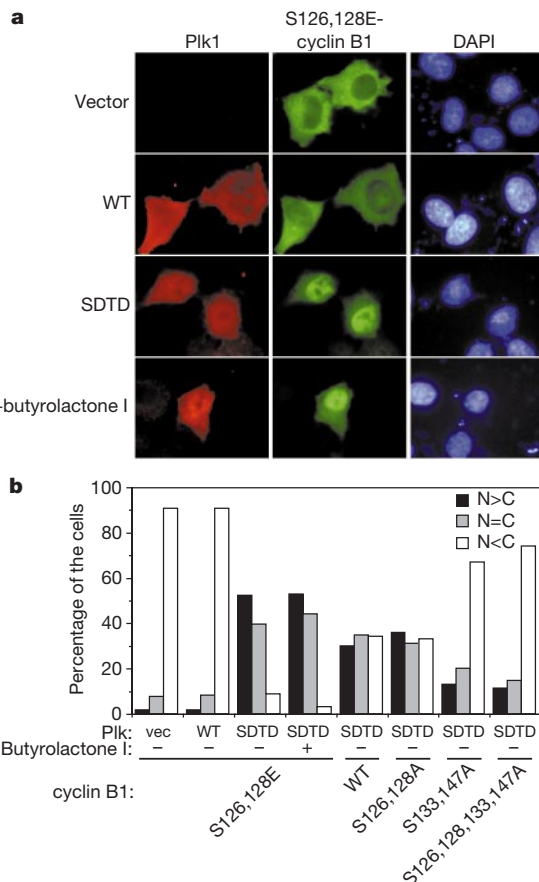


Figure 6 Constitutively active Plk1 induces nuclear localization of cyclin B1 in the absence of Cdc2 kinase activity. **a**, HeLa cells were co-transfected with a plasmid containing Myc–S126,128E-cyclin B1 and control vector, HA-tagged wild-type Plk1 or HA–SDTD-Plk1. After 15 h, cells were incubated with MG132 (50 μM) in the absence or presence of butyrolactone I (120 μM) for 6 h and then immunostained with anti-Myc and anti-HA antibodies. **b**, Cells were classified into three categories in terms of location of Myc–cyclin B1 as described in the legend of Fig. 1b. More than 200 cells were examined, and the percentages of each category are shown. Experiments were performed twice with similar results.

0.1 µg or 1 µg of His-tagged Plk1 or immunoprecipitates was mixed with substrates (casein or GST–cyclin B1, 0.1–1 µg), 50 µM ATP, 15 mM MgCl₂, 3 µCi of [γ -³²P]ATP and incubated for 20 min at 30 °C.

Purification of cyclin B1 kinase

De-jellied *Xenopus* eggs (50 ml) were homogenized in four volumes of buffer A (50 mM Tris pH 8.0, 25 mM 2-glycerophosphate, 10 mM EGTA, 10 mM MgCl₂, 0.1 mM NaF, 2 mM DTT, 1 mM Na₃VO₄, 2 mM PMSF) with 10 µg ml⁻¹ cytochalasin B and 20 µg ml⁻¹ aprotinin, centrifuged at 20,000g for 30 min and then 140,000g for 1 h, and was loaded onto a 50-ml Q-Sepharose Fast Flow (Pharmacia). Unadsorbed fractions were loaded onto a cyclin B1 affinity column, GST–S126,128,133,147A-cyclin B1 coupled to glutathione-Sepharose, equilibrated with buffer B (20 mM Tris pH 7.3, 12.5 mM 2-glycerophosphate, 2 mM EGTA, 2 mM MgCl₂, 2 mM DTT, 1 mM Na₃VO₄, 1 mM PMSF, 2 µg ml⁻¹ aprotinin). Proteins were eluted with a linear gradient to 1 M NaCl and assayed for the kinase activity toward GST–S126,128A-cyclin B1. Active fractions were loaded onto Hitrap–Heparin (Pharmacia), and eluted with a linear gradient to 1 M NaCl in buffer B. The active fractions were adjusted to 0.25 M NaCl, loaded onto oligo(dT)–cellulose (GibcoBRL) equilibrated with buffer B containing 0.25 M NaCl, and eluted stepwise with buffer B containing 0.4, 0.6, 0.8, 1.0 and 1.2 M NaCl.

Identification of proteins by peptide mass mapping

Samples were electrophoresed on a 9% polyacrylamide gel and transferred onto a PDVF membrane. The immobilized proteins were reduced, S-carboxymethylated, and digested *in situ* with Achromobacter protease I (a Lys-C)²⁹. Molecular mass analysis of Lys-C fragments was performed by spectrometry using a PerSeptive Biosystem Voyager-DE/RP³⁰. We identified proteins by comparing the molecular weights determined by MALDI-TOF/MS with theoretical peptide masses from the proteins registered in NCBItr (10 April 1999).

Anti-S147-phosphorylated cyclin B1 antibody (147P)

Anti-S147-phosphorylated cyclin B1 polyclonal antibody was produced in rabbits by immunizing them with a keyhole limpet haemocyanin-conjugated synthetic phosphopeptide corresponding to residues 141–154 of cyclin B1 (DLQCAF-phosphoS147-DVILAVN). The serum was affinity purified with the phosphopeptide- and the non-phosphopeptide (DLQCAFVILAVN)-conjugated cellulose. The purified 147P antibody was added to the cell extracts and the immunoprecipitates were washed three times and immunoblotted with anti-cyclin B1. In the immunoprecipitation with 147P from the immunoprecipitates with anti-cyclin B1, endogenous cyclin B1 was first immunoprecipitated with anti-cyclin B1, washed three times, and then released from the beads by incubating in the presence of 1% SDS. After the beads were discarded, supernatants were diluted 10 times and then immunoprecipitated with anti-147P. In the assay shown in Fig. 5e, bottom panel, anti-147P and agarose-conjugated anti-cyclin B1 or IgG were added to the extracts. Supernatants were immunoblotted with horseradish peroxidase-conjugated anti-rabbit immunoglobulins (Amersham) and anti-cyclin B1 to detect anti-147P and cyclin B1, respectively.

Received 31 October; accepted 15 December 2000.

- Nurse, P. Universal control mechanism regulating onset of M-phase. *Nature* **344**, 503–508 (1990).
- Hunt, T. Cyclins and their partners: from a simple idea to complicated reality. *Semin. Cell Biol.* **2**, 213–222 (1991).
- Morgan, D. O. Principles of CDK regulation. *Nature* **374**, 131–134 (1995).
- Pines, J. & Hunter, T. Human cyclins A and B1 are differentially located in the cell and undergo cell cycle-dependent nuclear transport. *J. Cell Biol.* **115**, 1–17 (1991).
- Ookata, K., Hisanaga, S., Okano, T., Tachibana, K. & Kishimoto, T. Relocation and distinct subcellular localization of p34cdc2–cyclin B complex at meiosis reinitiation in starfish oocytes. *EMBO J.* **11**, 1763–1772 (1992).
- Pines, J. & Hunter, T. The differential localization of human cyclins A and B is due to a cytoplasmic retention signal in cyclin B. *EMBO J.* **13**, 3772–3781 (1994).
- Heald, R., McLoughlin, M. & McKeon, F. Human wee1 maintains mitotic timing by protecting the nucleus from cytoplasmically activated Cdc2 kinase. *Cell* **74**, 463–474 (1993).
- Li, J., Meyer, A. N. & Donoghue, D. J. Nuclear localization of cyclin B1 mediates its biological activity and is regulated by phosphorylation. *Proc. Natl Acad. Sci. USA* **94**, 502–507 (1997).
- Jin, P., Hardy, S. & Morgan, D. O. Nuclear localization of cyclin B1 controls mitotic entry after DNA damage. *J. Cell Biol.* **141**, 875–885 (1998).
- Toyoshima, F., Moriguchi, T., Wada, A., Fukuda, M. & Nishida, E. Nuclear export of cyclin B1 and its possible role in the DNA damage-induced G2 checkpoint. *EMBO J.* **17**, 2728–2735 (1998).
- Yang, J. & Kornbluth, S. All aboard the cyclin train: subcellular trafficking of cyclins and their CDK partners. *Trends Cell Biol.* **9**, 207–210 (1999).
- Pines, J. Four-dimensional control of the cell cycle. *Nature Cell Biol.* **1**, E73–E79 (1999).
- Yang, J. *et al.* Control of cyclin B1 localization through regulated binding of the nuclear export factor CRM1. *Genes Dev.* **12**, 2131–2143 (1998).
- Hagting, A., Jackman, M., Simpson, K. & Pines, J. Translocation of cyclin B1 to the nucleus at prophase requires a phosphorylation-dependent nuclear import signal. *Curr. Biol.* **9**, 680–689 (1999).
- Hagting, A., Karlsson, C., Clute, P., Jackman, M. & Pines, J. MPF localization is controlled by nuclear export. *EMBO J.* **17**, 4127–4138 (1998).
- Kumagai, A. & Dunphy, W. G. Purification and molecular cloning of Plx1, a Cdc25-regulatory kinase from *Xenopus* egg extracts. *Science* **273**, 1377–1380 (1996).
- Glover, D. M., Hagan, I. M. & Tavares, A. A. Polo-like kinases: a team that plays throughout mitosis. *Genes Dev.* **12**, 3777–3787 (1998).
- Nigg, E. A. Polo-like kinases: positive regulators of cell division from start to finish. *Curr. Opin. Cell Biol.* **10**, 776–783 (1998).

- Jackman, M., Firth, M. & Pines, J. Human cyclins B1 and B2 are localized to strikingly different structures: B1 to microtubules, B2 primarily to the Golgi apparatus. *EMBO J.* **14**, 1646–1654 (1995).
- Nigg, E. A. Assembly–disassembly of the nuclear lamina. *Curr. Opin. Cell Biol.* **4**, 105–109 (1992).
- Izumi, T. & Maller, J. L. Phosphorylation of *Xenopus* cyclins B1 and B2 is not required for cell cycle transitions. *Mol. Cell. Biol.* **11**, 3860–3867 (1991).
- Li, J., Meyer, A. N. & Donoghue, D. J. Requirement for phosphorylation of cyclin B1 for *Xenopus* oocyte maturation. *Mol. Biol. Cell* **6**, 1111–1124 (1995).
- Borgne, A., Ostvold, A. C., Flament, S. & Meijer, L. Intra-M phase-promoting factor phosphorylation of cyclin B at the prophase/metaphase transition. *J. Biol. Chem.* **274**, 11977–11986 (1999).
- Golsteyn, R. M., Mundt, K. E., Fry, A. M. & Nigg, E. A. Cell cycle regulation of the activity and subcellular localization of Plk1, a human protein kinase implicated in mitotic spindle function. *J. Cell Biol.* **129**, 1617–1628 (1995).
- Hamanaka, R. *et al.* Polo-like kinase is a cell cycle-regulated kinase activated during mitosis. *J. Biol. Chem.* **270**, 21086–21091 (1995).
- Qian, Y. W., Erikson, E. & Maller, J. L. Mitotic effects of a constitutively active mutant of the *Xenopus* Polo-like kinase Plx1. *Mol. Cell. Biol.* **19**, 8625–8632 (1999).
- Qian, Y. W., Erikson, E., Li, C. & Maller, J. L. Activated polo-like kinase Plx1 is required at multiple points during mitosis in *Xenopus laevis*. *Mol. Cell. Biol.* **18**, 4262–4271 (1998).
- Fukuda, M. *et al.* CRM1 is responsible for intracellular transport mediated by the nuclear export signal. *Nature* **390**, 308–311 (1997).
- Iwamatsu, A. S-carboxymethylation of proteins transferred onto polyvinylidene difluoride membranes followed by *in situ* protease digestion and amino acid microsequencing. *Electrophoresis* **13**, 142–147 (1992).
- Jensen, O. N., Podtelejnikov, A. & Mann, M. Delayed extraction improves specificity in database searches by matrix-assisted laser desorption/ionization peptide maps. *Rapid Commun. Mass Spectrom.* **10**, 1371–1378 (1996).

Acknowledgements

We are grateful to K. Todokoro for a mouse Plk1 complementary DNA and to H. Ellinger-Ziegelbauer for advice on Plk1 assays. We thank M. Yanagida for helpful discussion. This work was supported by grants from the Ministry of Education, Science and Culture of Japan (E.N.).

Correspondence and requests for materials should be addressed to E.N. (e-mail: L50174@sakura.kudpc.kyoto-u.ac.jp).

The S-locus receptor kinase is inhibited by thioredoxins and activated by pollen coat proteins

Didier Cabrillac, J. Mark Cock, Christian Dumas & Thierry Gaudé

Reproduction et Développement des Plantes, UMR 5667 CNRS-INRA-ENSL-UCBL Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon Cedex 07, France

The self-incompatibility response in *Brassica* allows recognition and rejection of self-pollen by the stigmatic papillae. The transmembrane S-locus receptor kinase (SRK), a member of the receptor-like kinase superfamily in plants, mediates recognition of self-pollen on the female side¹, whereas the S-locus cysteine-rich protein (SCR) is the male component of the self-incompatibility response². SCR is presumably located in the pollen coat, and is thought to be the SRK ligand^{2,3}. Although many receptor-like kinases have been isolated in plants, the mechanisms of signal transduction mediated by these molecules remain largely unknown. Here we show that SRK is phosphorylated *in vivo* within one hour of self-pollination. We also show that, *in vitro*, autophosphorylation of SRK is prevented by the stigma thioredoxin THL1 in the absence of a ligand. This inhibition is released in a haplotype-specific manner by the addition of pollen coat proteins. Our data indicate that SRK is inhibited by thioredoxins and activated by pollen coat proteins.

To investigate the phosphorylation status of SRK during the self-incompatibility reaction, we labelled pistils from self-incompatible S₃ homozygous *Brassica oleracea* plants with [³²P]orthophosphate, and monitored phosphorylation of the receptor during a self-pollination time course. Physiological studies have shown that

pollen hydrates 30–90 min after its deposition on the stigmatic papillae⁴. SRK was phosphorylated about 60 min after self-pollination but, as with pollen hydration, the time course of SRK phosphorylation varied between 45–60 min depending on the experimental conditions (Fig. 1a). No phosphorylation of SRK was detected up to 90 min after compatible pollination (data not shown).

Only a fraction of the SRK protein was phosphorylated *in vivo* (compare Fig. 1a and b, left panel). Stigmatic papillae are able to discriminate between a self- and a cross-pollen deposited on the same cell⁵, indicating that the self-incompatibility reaction is highly localized at the subcellular level. Our data are consistent with SRK being activated only in regions of the stigma papillar cells that contact the pollen grains.

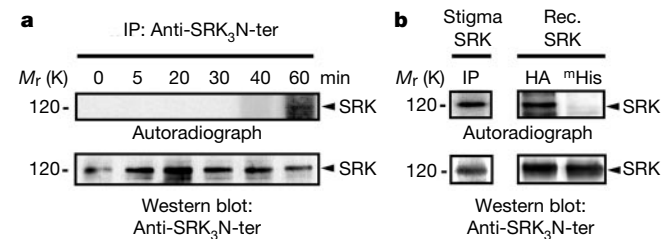


Figure 1 SRK is phosphorylated 60 min after self-pollination *in vivo*, but is constitutively active *in vitro*. **a**, Time course of SRK phosphorylation *in vivo*. SRK was immunoprecipitated from extracts of radiolabelled stigmas and detected either by autoradiography (top) or immunodetection (bottom). **b**, SRK is constitutively active *in vitro*. *In vitro* phosphorylation reactions were carried out with stigma microsomes containing SRK (left) and with microsomes from insect cells expressing either functional, HA-tagged SRK (HA) or kinase-inactive, His-tagged SRK (³⁵His) (right). Rec. SRK, recombinant SRK.

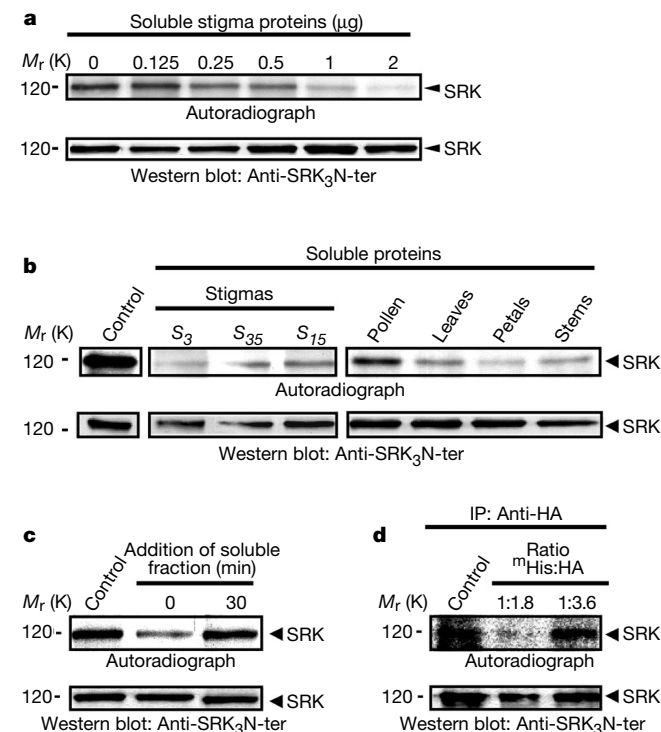


Figure 2 Phosphorylation of SRK *in vitro* is inhibited by a soluble protein. **a**, Dose-dependent inhibition of SRK phosphorylation by a soluble stigma extract. **b**, The soluble inhibitor of SRK phosphorylation is neither haplotype-specific nor tissue-specific. **c**, SRK is not dephosphorylated by soluble stigma proteins added 30 min after the beginning of the phosphorylation reaction. **d**, Kinase-inactive SRK (³⁵His) competes with functional, HA-tagged SRK (HA) for the binding of the inhibitor. In **b–d**, 0.6 µg of soluble protein was added; no extract was added to controls.

In contrast to these results obtained *in vivo*, recombinant SRK, expressed in a baculovirus/insect cell expression system, was constitutively active *in vitro*⁶ (Fig. 1b). This difference was not caused by overexpression, because SRK was also constitutively active in microsomes isolated from stigmas (Fig. 1b). One possible explanation for these conflicting results was the existence of an inhibitory factor *in vivo* that was absent from the microsome fraction used in the *in vitro* assay. To test this hypothesis, we added different amounts of a soluble stigma extract to recombinant SRK in an *in vitro* phosphorylation assay. Under these conditions, SRK phosphorylation decreased in a dose-dependant manner (Fig. 2a), suggesting that the soluble stigma extract contained a negative regulator of SRK phosphorylation. A similar result was obtained when stigma-extracted SRK was used instead of recombinant SRK (data not shown).

The inhibitory effect of the soluble stigma fraction was lost when it was boiled or treated with trypsin but not after treatment with denatured trypsin, indicating that the active component was a protein (data not shown). We tested soluble extracts from stigmas of *Brassica* lines carrying different *S* haplotypes and extracts from different tissues. All extracts had an inhibitory effect although the various tissue extracts varied in their ability to inhibit SRK activity, despite the fact that equal amounts of protein were added (Fig. 2b). These experiments showed that the inhibitory activity was present in a wide range of tissues, albeit at different concentrations, and was not *S*-haplotype-specific.

Although phosphatase inhibitors were used in the phosphorylation buffer, it was nonetheless possible that the observed activity might be the result of a non-inactivated phosphatase. To test this hypothesis, we added the soluble stigma extract either immediately

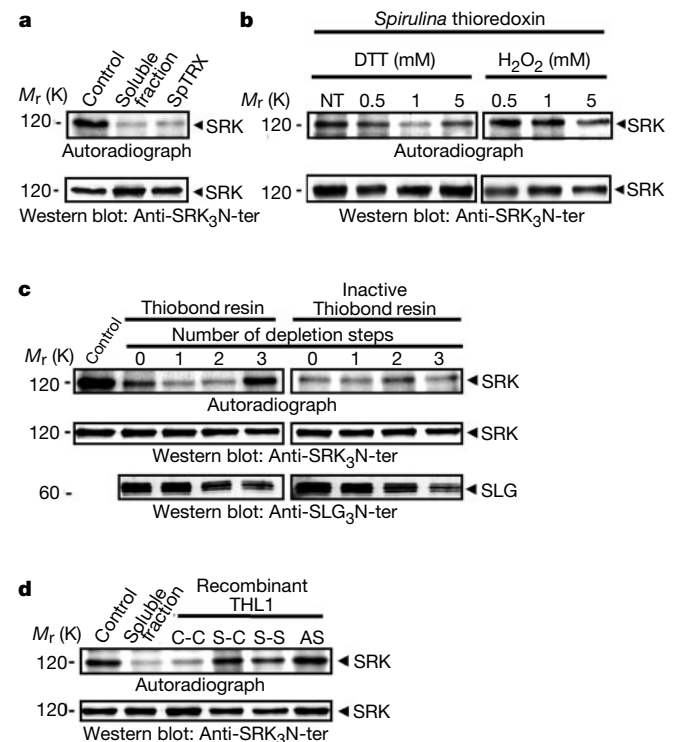


Figure 3 Thioredoxin inhibits SRK phosphorylation *in vitro*. **a**, *Spirulina* thioredoxin (SpTRX, 20 ng) has the same inhibitory effect as soluble stigma proteins (0.6 µg). **b**, The inhibitory activity of *Spirulina* thioredoxin is redox dependent. SpTRX (0.6 µg) was pretreated with redox compounds as indicated. NT, not treated. **c**, Depletion of thioredoxin from the stigma extract by functional Thiobond resin abolishes the inhibitory activity of the extract. **d**, *In vitro* phosphorylation of SRK is inhibited by 15 ng of *Brassica* THL1 (C-C), but not by the inactive mutant forms (S-C and S-S) nor by an extract from an antisense control (AS).

or 30 min after the beginning of the phosphorylation reaction (at which point SRK is already phosphorylated; see Fig. 1b), and incubated the reactions for an additional 30 min. Inhibition of SRK activity occurred only if the soluble extract was added at the beginning of the reaction (Fig. 2c), indicating that it prevents SRK phosphorylation rather than dephosphorylates this protein.

To test whether the decrease in SRK activity was due to the inhibitor binding directly to SRK, we performed competition experiments in an *in vitro* phosphorylation reaction. To the phosphorylation reaction medium, we added an excess of an inactive recombinant form of SRK (³⁵S-His), together with the functional recombinant haemagglutinin (HA)-tagged SRK, and soluble stigma extract. When the ³⁵S-His competitor was added in a 3.6-fold excess, SRK-HA kinase activity was not inhibited by the stigma extract (Fig. 2d). This experiment indicated that the inhibitor bound to both the wild-type and mutant forms of SRK in a stoichiometric and phosphorylation-independent manner.

Arm repeat containing 1 (ARC1) interacts specifically with the cytoplasmic domain of SRK⁷, but it was unlikely to correspond to the inhibitor because it interacts only with the phosphorylated form of SRK and is expressed specifically in stigmas. Furthermore, reduced levels of ARC1 in transgenic plants are correlated with an attenuation of the self-incompatibility response⁸. Two other proteins, thioredoxin-h-like (THL)-1 and THL2, were better candidates because they bind the cytosolic domain of SRK in a phosphorylation-independent manner⁹. Moreover, in mammals, thioredoxin (TRX) binds and regulates the activity of cytosolic kinases such as apoptosis signal-regulating kinase 1 (ASK1)¹⁰ and p38 MAP kinase¹¹.

We looked at the possible regulatory effect of thioredoxin on SRK activity by adding purified *Spirulina* thioredoxin (SpTRX) to SRK phosphorylation assays. SpTRX mimicked the effect of the stigma extract (Fig. 3a). Dithiothreitol (DTT), which can be used by thioredoxins as an electron donor, is necessary for the inhibitory effect of TRX on ASK1 (ref. 10). SpTRX lost its inhibitory activity when DTT was omitted from the buffer or when SpTRX was treated with H₂O₂ (Fig. 3b). SRK phosphorylation was not affected by these treatments when SpTRX was omitted (data not shown). The effect of DTT was dose dependant, but excess DTT (5 mM) suppressed the action of SpTRX, possibly by preventing disulphide bridge formation between the enzyme and SRK.

Specific depletion of thioredoxins from the stigma extract by

repeated incubation with Thiobond resin abolished the inhibitory effect of the soluble stigma extract (Fig. 3c, top). Inactivated resin did not deplete the inhibitory activity. The S-locus glycoprotein (SLG, a soluble glycoprotein expressed in stigmas¹²) was not significantly depleted from the extract, indicating that loss of inhibitory activity was not due to a general depletion of protein (Fig. 3c, bottom). Together, these results indicated that the inhibitory effect of the stigma extract was caused by the activity of a thioredoxin.

We confirmed that THL1 can inhibit SRK activity by adding a recombinant form of THL1, expressed in *Escherichia coli*, to phosphorylation reactions *in vitro* (Fig. 3d). To get further insight into the mechanism of action of this enzyme, we produced two mutated forms of THL1 in which one or both of the catalytic cysteines were replaced by serines (THL1^{C45S} and THL1^{C45S/C48S}). These mutations suppress thioredoxin activity of the enzyme by impairing its catalytic mechanism¹³. The two mutant forms of THL1 did not inhibit SRK phosphorylation *in vitro*, suggesting that the catalytic activity of THL1 is necessary for it to act as an inhibitor.

In *Brassica*, the male component of the self-incompatibility response is a pollen coat protein (PCP)³. When PCPs extracted from S₃ or S₁₅ pollen were added to recombinant SRK in the presence of the inhibitor, SRK was phosphorylated in a dose-dependant and haplotype-specific manner (Fig. 4a). This result indicates that PCPs can activate SRK in the presence of the inhibitor. In animal receptor kinases, ligands activate their receptor by inducing their oligomerization and transphosphorylation¹⁴. This activation can be mimicked by the addition of bivalent antibodies that create bridges between receptor molecules¹⁵. Notably, in the presence of the antibody anti-SRK₃N-ter, recombinant SRK was phosphorylated in a dose-dependant and saturable manner (as expected for an antigen-antibody interaction), mimicking the action of PCPs (Fig. 4b). Similar results were obtained with stigma-extracted SRK (data not shown). By analogy with the studies carried out on animal receptors, these data indicate that the activation of SRK by PCPs may involve receptor oligomerization. This is consistent with a previous study that showed that transphosphorylation occurs between SRK molecules⁶. To our knowledge, this is the first demonstration that a plant receptor-like kinase is phosphorylated in response to a specific stimulus.

In summary, our data strongly support a model in which THL1 (and THL2) interacts in a reversible manner with SRK in stigmas and prevents spontaneous activation of the self-incompatibility signalling pathway. In this model, SRK thioredoxin inhibition is released by PCPs after self-pollination. In the absence of this regulation, it is probable that the self-incompatibility system would be constitutively active, blocking both self-pollination and cross-pollination and thus preventing seed set. □

Methods

Extraction and analysis of stigma and recombinant SRK

Insect cell microsomes expressing recombinant, HA-tagged and kinase-inactive (K553R), 6× His-tagged SRK₃ proteins have been described⁶. We prepared crude stigma extracts, plant microsomes and soluble fractions from *B. oleracea* lines as described¹⁶, except that for microsome preparation we used the following extraction buffer: 50 mM HEPES pH 7.4, 0.5 M sucrose, 1 mM DTT, 10 mM EDTA, 10 mM NaF, 5 mM β-glycerophosphate plus a protease inhibitor cocktail (Boehringer). After ultracentrifugation, the microsome pellet was resuspended in 50 mM HEPES pH 7.4 and stored, with soluble fraction (supernatant), at -80 °C before use. Proteins were separated on a 7.5% SDS-PAGE gel, electroblotted, and analysed by autoradiography or immunodetected with anti-SRK₃N-ter (monoclonal antibody 85-36-71) as described^{16,17}.

Immunoprecipitation of SRK

We immunoprecipitated SRK in 1 ml of TBST buffer (20 mM Tris-HCl pH 8, 150 mM NaCl, 0.05% w/v Tween 20), complemented with 10 mM EDTA, a protease inhibitor cocktail (Boehringer) and phosphatase inhibitors (5 mM orthovanadate, 10 mM β-glycerophosphate, 10 mM NaF). Immunoprecipitation was carried out with monoclonal anti-SRK₃N-ter (64-32-30)¹⁷ or monoclonal anti-HA (12CA5, Boehringer) and Protein G/Agarose (Tebu) according to the instructions of the manufacturer.

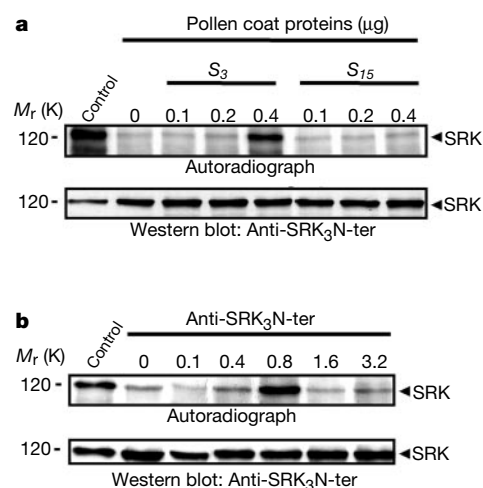


Figure 4 SRK is phosphorylated after adding an agonist in the presence of thioredoxin. **a**, After inhibition of SRK phosphorylation in the presence of soluble stigma extract, addition of S₃, but not S₁₅, PCPs induces SRK kinase activity. **b**, Under the same conditions, addition of a bivalent antibody (anti-SRK₃N-ter) also induces phosphorylation of SRK.

In vivo radioisotope labelling

Pistils from *S₃* homozygous *Brassica oleracea* plants were radiolabelled at room temperature in a microplate well in 10 mM MES pH 6, complemented with 25 μ Ci per pistil of [³²P]orthophosphate (ICN; specific activity 285 Ci mg⁻¹, 1.25 μ Ci μ l⁻¹). After a 4-h incubation, the incubation buffer was replaced by 10 mM MES buffer pH 6 and pistils were self-pollinated. We immunoprecipitated SRK from crude protein extracts as above.

Phosphorylation in vitro

Typically, stigma microsomes or insect cell microsomes expressing SRK (10 μ g protein) were labelled for 30 min at room temperature using 5 μ Ci of [γ -³²P]ATP (4 μ Ci μ l⁻¹, specific activity 3,000 Ci mmol⁻¹; Amersham-Pharmacia), as described⁶. For stigma microsomes, ATP was removed at the end of the reaction by desalting on a hand-made column containing 600 μ l of P6DG Gel (Biorad) equilibrated in 20 mM Tris-HCl pH 8, 2.5% SDS. SRK was immunoprecipitated before SDS-PAGE analysis. PCPs added to phosphorylation reactions were prepared as described¹⁸.

Bacterial and recombinant stigma thioredoxins

Spirulina thioredoxin (Sigma) was resuspended in microsome extraction buffer and incubated for 30 min on ice. For the redox experiments (Fig. 3b), DTT was omitted, added at different concentrations, or replaced by H₂O₂. *THL1* was amplified from a *B. oleracea* stigma complementary DNA library¹⁹ using oligonucleotides TH1 (5'-GGAGGGATCC ATGGCGCAACAGCAGAAGTG-3') and TH2 (5'-CTCAGGATCCGTCGGCACCAC AACACAAAGG-3') by polymerase chain reaction (PCR). Mutant forms of THL1 (THL1^{C45S} and THL1^{C45S/C48S}) were obtained using PCR mutagenesis with oligonucleotides TH3 (5'-CACAGCAGTGTGGTGGCCACCTAGCCGTTTCATTG-3'), TH4 (5'-CACAGCAGTGTGGAGCCACCTAGCCGTTTCATTG-3') and TH5 (5'-AAGAAAGA AGGATAATCAAGCGGCAGCAACT-3'). In Fig. 3, wild-type THL1, THL1^{C45S} and THL1^{C45S/C48S} are referred to as C-C, S-C and S-S, respectively. The three *THL1* constructs were verified by DNA sequencing and inserted into pQE30 (Qiagen) for expression in *E. coli*. We purified the expressed proteins on a Ni-NTA affinity column under native conditions using the Ni-NTA spin Kit (Qiagen) according to the manufacturer's instructions. Imidazole was removed from eluates by desalting as above, except that the column was equilibrated in microsome extraction buffer.

Thioredoxin depletion on Thiobond resin

One bed volume of Thiobond resin (Invitrogen) was reduced in two bed volumes of 50 mM HEPES pH 7.4, 20 mM β -mercaptoethanol for 30 min at room temperature. Inactive resin was produced by incubating one bed volume of reduced resin in two bed volumes of 20 mM 4-vinyl pyridine²⁰ after reduction. Active and inactive resins were washed and resuspended in 50 mM HEPES pH 7.4 before use. We carried out thioredoxin depletion by active and alkylated resins for 30 min at room temperature under shaking, using 4 μ l of bed resin per μ g protein. After incubation, unbound proteins were recovered by decantation of the resin and were incubated two more times with fresh resin. We tested an aliquot of each fraction in an *in vitro* phosphorylation assay. Nonspecific adsorption of protein on the beads was monitored by immunodetection of SLG with anti-SLG₃N-ter¹⁶.

Received 27 October; accepted 8 December 2000.

1. Takasaki, T. *et al.* The S receptor kinase determines self-incompatibility in *Brassica* stigma. *Nature* **403**, 913–916 (2000).
2. Schopfer, C. R., Nasrallah, M. E. & Nasrallah, J. B. The male determinant of self-incompatibility in *Brassica*. *Science* **286**, 1697–1700 (1999).
3. Stephenson, A. G. *et al.* The male determinant in self-incompatibility in *Brassica oleracea* is located in the pollen coating. *Plant J.* **12**, 1351–1359 (1997).
4. Dickinson, H. G. Dry stigmas, water and self-incompatibility in *Brassica*. *Sex Plant Reprod.* **8**, 1–10 (1995).
5. Sarker, R. H., Elleman, C. J. & Dickinson, H. G. Control of pollen hydration in *Brassica* requires continued protein synthesis and glycosylation is necessary for intraspecific incompatibility. *Proc. Natl Acad. Sci. USA* **85**, 4340–4344 (1988).
6. Giranton, J. L., Dumas, C., Cock, J. M. & Gaude, T. The integral membrane S-locus receptor kinase of *Brassica* has serine/threonine kinase activity in a membranous environment and spontaneously forms oligomers in planta. *Proc. Natl Acad. Sci. USA* **97**, 3759–3764 (2000).
7. Gu, T., Mazzurco, M., Sulaman, W., Matias, D. D. & Goring, D. R. Binding of an arm repeat protein to the kinase domain of the S-locus receptor kinase. *Proc. Natl Acad. Sci. USA* **95**, 382–387 (1998).
8. Stone, S. L., Arnoldo, M. & Goring, D. R. A breakdown of *Brassica* self-incompatibility in *ARC1* antisense transgenic plants. *Science* **286**, 1729–1731 (1999).
9. Bower, M. S. *et al.* Two members of the thioredoxin-h family interact with the kinase domain of a *Brassica* S locus receptor kinase. *Plant Cell* **8**, 1641–1650 (1996).
10. Saitoh, M. *et al.* Mammalian thioredoxin is a direct inhibitor of apoptosis signal-regulating kinase (ASK). *J. EMBO J.* **17**, 2596–2606 (1998).
11. Hashimoto, S. *et al.* Thioredoxin negatively regulates p38 MAP kinase activation and IL-6 production by tumor necrosis factor- α . *Biochem. Biophys. Res. Commun.* **258**, 443–447 (1999).
12. Kandasamy, M. K., Paolillo, D. J., Faraday, C. D., Nasrallah, J. B. & Nasrallah, M. E. The S-locus specific glycoproteins of *Brassica* accumulate in the cell wall of developing stigma papillae. *Dev. Biol.* **134**, 462–472 (1989).
13. Meyer, Y., Verdoucq, L. & Vignols, F. Plant thioredoxins and glutaredoxins: identity and putative roles. *Trends Plant Sci.* **4**, 388–394 (1999).
14. Heldin, C. H. Dimerization of cell surface receptors in signal transduction. *Cell* **80**, 213–223 (1995).
15. Spaargaren, M., Defize, L. H., Boonstra, J. & de Laat, S. W. Antibody-induced dimerization activates the epidermal growth factor receptor tyrosine kinase. *J. Biol. Chem.* **266**, 1733–1739 (1991).
16. Delorme, V. *et al.* Characterization of the S locus genes, *SLG* and *SRK*, of the *Brassica* *S₃* haplotype: identification of a membrane-localized protein encoded by the S locus receptor kinase gene. *Plant J.* **7**, 429–440 (1995).

17. Gaude, T. *et al.* Expression of a self-incompatibility gene in a self-compatible line of *Brassica oleracea*. *Plant Cell* **5**, 75–86 (1993).
18. Doughty, J., Hedderson, F., Mc Cubbin, A. & Dickinson, H. G. Interaction between a coating-borne peptide of the *Bassica* pollen grain and S (incompatibility)-locus linked stigmatic glycoproteins. *Proc. Natl Acad. Sci. USA* **90**, 467–471 (1993).
19. Giranton, J. L., Ariza, M. J., Dumas, C., Cock, J. M. & Gaude, T. The S locus receptor kinase gene encodes a soluble glycoprotein corresponding to the SKR extracellular domain in *Brassica oleracea*. *Plant J.* **8**, 827–834 (1995).
20. Flannery, A. V., Beynon, R. J. & Bond, J. S. in *Proteolytic Enzymes, a Practical Approach* (eds Beynon, R. J. & Bond, J. S.) 145–162 (IRL, Oxford, 1989).

Acknowledgements

We thank J.-L. Giranton for production of baculovirus-insect cell microsomes and A.-M. Thierry for her technical assistance. J.M.C. is a member of the Institut National de la Recherche Agronomique and T.G. is a member of the Centre National de la Recherche Scientifique.

Correspondence and requests for materials should be addressed to T.G. (e-mail: thierry.gaude@ens-lyon.fr). The GenBank accession number of BoTHL1 is AF273844.

Conformational diversity in a yeast prion dictates its seeding specificity

Peter Chien & Jonathan S. Weissman

Howard Hughes Medical Institute, Department of Cellular and Molecular Pharmacology, Graduate Group in Biophysics, University of California-San Francisco, San Francisco, California 94143-0450, USA

A perplexing feature of prion-based inheritance is that prions composed of the same polypeptide can evoke different phenotypes (such as distribution of brain lesions), even when propagated in genetically identical hosts^{1–3}. The molecular basis of this strain diversity and the relationship between strains and barriers limiting transmission between species remain unclear. We have used the yeast prion phenomenon [*PSI*⁺]⁴ to investigate these issues and examine the role that conformational differences^{1,2,5} may have in prion strains. We have made a chimaeric fusion between the prion domains of two species (*Saccharomyces cerevisiae* and *Candida albicans*) of Sup35, the protein responsible for [*PSI*⁺]^{4,6}. Here we report that this chimaera forms alternate prion strains *in vivo* when initiated by transient overexpression of different Sup35 species. Similarly, *in vitro* the purified chimaera, when seeded with different species of Sup35 fibres, establishes and propagates distinct amyloid conformations. These fibre conformations dictate amyloid seeding specificity: a chimaera seeded by *S. cerevisiae* fibres efficiently catalyses conversion of *S. cerevisiae* Sup35 but not of *C. albicans* Sup35, and vice versa. These and other considerations^{1,2,5,7,8} argue that heritable prion strains result from self-propagating conformational differences within the prion protein itself. Moreover, these conformational differences seem to act in concert with the primary structure to determine a prion's propensity for transmission across a species barrier.

The [*PSI*⁺] prion element of the yeast *S. cerevisiae* is transmitted through self-propagating aggregates of Sup35, a component of the eukaryotic translation release factor⁹. Loss of soluble Sup35 in [*PSI*⁺] cells leads to a heritable suppression of nonsense mutations^{4,6}, which can be readily monitored using a nonsense mutation in the *ADE1* gene (*ade1-14*)¹⁰. The suppression of *ade1-14* allows growth on media lacking adenine, and prevents the accumulation of metabolic intermediates that cause [*psi*⁻] colonies to appear red on low adenine media. [*PSI*⁺] propagation is mediated

by the (glutamine/asparagine)-rich amino-terminal prion-determining domain (PrD)^{11,12}. Transient overexpression of this domain causes formation of *de novo* Sup35 prion aggregates *in vivo*¹³ and purified PrD forms self-propagating amyloid fibres *in vitro*^{14,15}.

The prion-forming ability of Sup35 is strongly conserved across budding yeast (Sacchomycetes), but a barrier limits seeding between prion domains from different species^{16–18}. For example, we previously observed that overexpression of *C. albicans* Sup35 PrD (CA) did not induce conversion of *S. cerevisiae* Sup35 to [PSI], and vice versa¹⁷. We found that a short region (such as residues 1–39) of *S. cerevisiae* PrD (SC) could confer susceptibility to seeding by *S. cerevisiae* prion when placed in CA. However, this region alone does not support prion formation¹⁹ (our unpublished data).

To examine further the relationship between primary structure and species specificity, we determined whether the extreme N terminus of CA was similarly required for susceptibility to

C. albicans prion. These studies used a previously described chimaeric PrD¹⁷ (CHIM), comprising residues 1–39 of *S. cerevisiae* PrD and residues 40–140 from the *C. albicans* PrD. We monitored specificity of prion formation using a set of isogenic *S. cerevisiae* yeast strains in which the sole chromosomal copy of SUP35 consisted of SC, CA or CHIM PrDs fused to a highly charged middle domain (M) followed by the carboxy-terminal translation termination domain (EF), designated SC-EF, CA-EF and CHIM-EF, respectively. In these strains, the nonsense-suppression phenotype depended on the state (aggregated or soluble) of the relevant PrD fusions. To test for cross-species seeding, we used inducible plasmids containing the appropriate prion domain fused to green fluorescent protein (GFP). Here, the nonsense suppression phenotype, after transient overexpression, reports on the ability of the 'inducer' species to convert the chromosomal Sup35.

Naively, if the extreme N terminus of CA were required for interaction with CA, we would expect that CHIM would seed only itself and SC, but not CA. Instead, we found that CHIM was a promiscuous prion. Consistent with earlier results¹⁷, overexpression of CHIM induced formation of [PSI] in SC-EF (Fig. 1a), whereas overexpression of CA did not; however, transient CHIM overexpression efficiently induced CA aggregation (Fig. 1b), despite the presence of the extreme N terminus derived from *S. cerevisiae*. Overexpression of SC, CA or CHIM in a CHIM-EF background also initiated a [PSI]-like state (Fig. 1c), which, like all characterized yeast prions, was readily cured by growth on low concentrations (5 mM) of guanidine hydrochloride (Fig. 1c, inset)^{17,20–22}. Thus, CHIM is itself a functional prion domain that bridges the species barrier between *C. albicans* and *S. cerevisiae*.

Although SC or CA overexpression induced prion formation of CHIM-EF with comparable efficiency, the resulting [PSI]-like phenotypes were markedly different. After selecting prion convertants on medium lacking adenine, we examined their strain phenotypes by replating them on low adenine medium. This procedure revealed distinct colour phenotypes depending on the inducer species (Fig. 1d). Quantitative analysis (Fig. 1e) indicated that about 80% of the primary CHIM-EF prion inductants initiated by SC were either weak (pink) or unstable (sectoring); in contrast, over 90% of CA-induced colonies had strong (white) phenotypes.

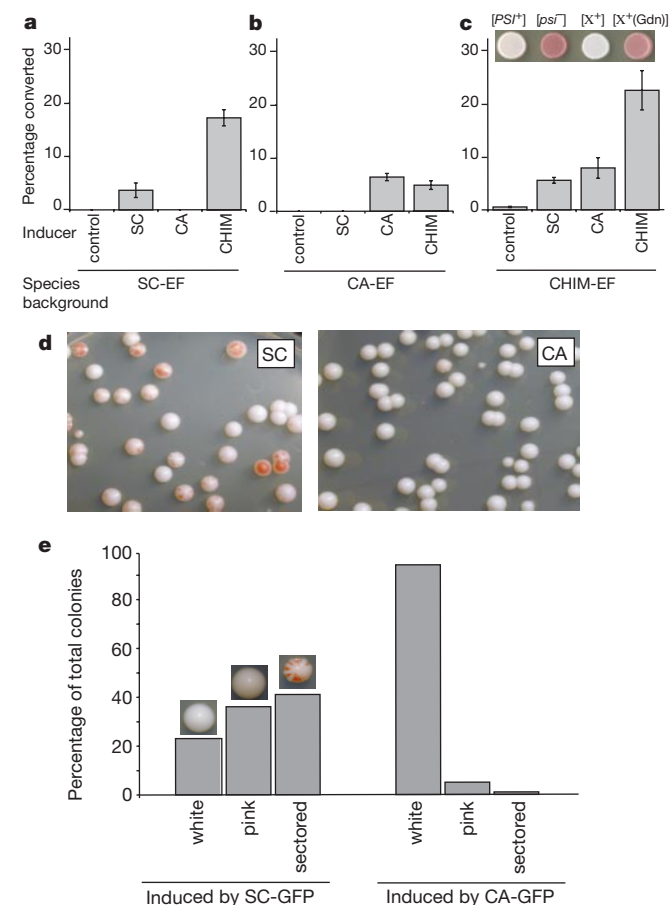


Figure 1 Characterization of CHIM prion formation *in vivo*. **a–c**, Induction experiments carried out in an SC-EF (**a**), CA-EF (**b**) or CHIM-EF (**c**) background, using the indicated species of inducer prion. Percentages of Ade⁺ colonies are shown. Control denotes empty vector. Increased induction of SC-EF by CHIM compared with by SC is probably caused by a higher rate of *de novo* aggregation of the CHIM prion domain, as observed by fluorescence microscopy (our unpublished observations) and spontaneous prion formation (**c**). Inset, induced chimaera prion [X⁺] shows curing after passage on 5 mM guanidine-HCl-containing medium [X⁺(Gdn)]; also shown are [PSI⁺] and [psi⁻] yeast for comparison. **d**, Examples of phenotypes of independently induced prion states of CHIM-EF with SC (left) or CA (right) inducers. CHIM-EF was induced by transient overexpression of the indicated inducer Sup35 species, and prion convertants were selected on minus Ade medium. For each inducer species, we pooled several convertants, back diluted and plated on low Ade medium to reveal strain phenotypes. **e**, Quantitative analysis of CHIM strains generated in **c**. Colonies were grouped according to colour (white, pink, sectoring). A representative picture of each colony type is shown above the appropriate bin.

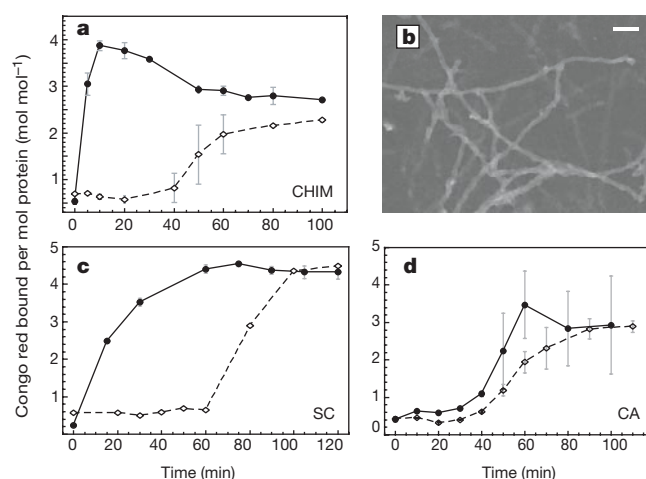


Figure 2 CHIM forms self-seeded amyloid fibres *in vitro*. **a**, Congo red binding assays show that purified CHIM monomer spontaneously forms amyloid with a lag phase (diamonds), which is abolished on addition of 3% (mol mol⁻¹) preformed fibres (circles). Abscissa represents moles of Congo red bound per mole of protein. **b**, Electron microscopy indicates that CHIM forms typical amyloid fibres. Scale bar, 100 nm. **c**, CHIM fibres from **a** catalyse SC polymerization (circles) compared with unseeded reactions (diamonds). **d**, Same CHIM fibres have only a modest effect on CA polymerization (circles) compared with unseeded CA reactions (diamonds).

Similar strain differences occur sporadically in new prion inducants of *S. cerevisiae*^{3,23} and other species of Sup35 (refs 16, 17). Unlike previous work on spontaneously generated strains, however, we can now control strain formation by changing the inducing species, thereby facilitating efforts to study this phenomenon *in vitro*.

We first established that purified CHIM could adopt an amyloid conformation by using a well-characterized *in vitro* system^{14,15}, assayed with the amyloid-specific dye Congo red¹². Similar to wild-type SC polymerization, spontaneous conversion of CHIM was accompanied by a lag phase of about 50 min, which was eliminated by the addition of a small amount (3% mol mol⁻¹) of the fibres formed at the reaction endpoint (Fig. 2a). As in previous studies²⁴, we found that the kinetics of Congo red binding mirrored

formation of fibres observed by electron microscopy (Fig. 2b), and increased binding to thioflavin-T and resistance to solubilization by SDS in the absence of boiling (data not shown). The spontaneously formed CHIM fibres readily seeded SC (Fig. 2c); however, in contrast to our *in vivo* results, little seeding of CA polymerization was observed (Fig. 2d). No such difference between *in vivo* and *in vitro* seeding had been previously reported in over a dozen experiments with different species and mutants of Sup35 (refs 12, 17, 19, 25, 26; and see below).

We next established that induction of CHIM by SC or CA could lead to different fibre conformations. Akin to *in vivo* results, both SC fibres and CA fibres efficiently seeded conversion of CHIM protein (Fig. 3a). Polymerization of CHIM monomer initiated by SC fibres (CHIM[SC]) led to a monotonic increase in apparent Congo red binding; however, CHIM monomer seeded by CA fibres (CHIM[CA]) showed a fast rise and steady decline in apparent Congo red binding. This profile was highly reminiscent of CA polymerization seeded by CA (Fig. 3a, inset). The loss of apparent Congo red binding is associated with formation of higher-order fibre aggregates and is not due to depolymerization, as electron microscopy and thioflavin-T binding confirmed the persistence of amyloid (data not shown). Moreover, brief sonication of the final material (data not shown) or increasing shear during conversion, as we did in subsequent CA experiments (Fig. 4b), restored apparent Congo red binding. Thus, the decrease in dye binding results from a prion conformation favouring formation of higher-order fibre aggregates; this property is common to CA and CHIM[CA] fibres but not CHIM[SC] fibres.

The two distinct CHIM fibre conformations propagated stably (Fig. 3b), as the characteristic differences in apparent Congo red binding were also observed in second-generation fibres in which CHIM was seeded by CHIM[SC] or CHIM[CA], denoted CHIM[[SC]] and CHIM[[CA]], respectively. Limited proteolysis and detection by western blot analysis²⁴ confirmed the existence of

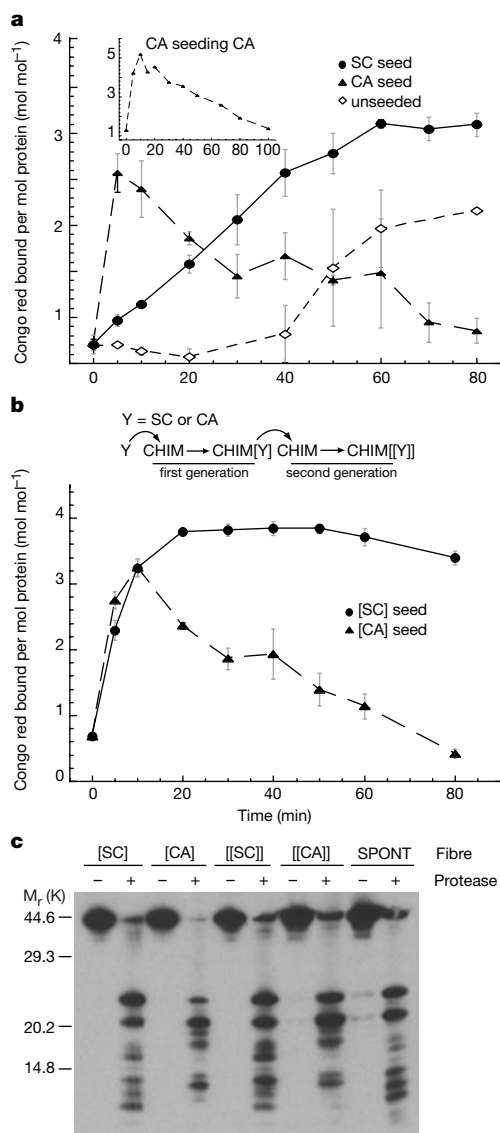


Figure 3 Generation and propagation of distinct CHIM fibre conformations. **a**, Polymerization of CHIM in the absence (diamonds) or presence of 3% SC fibres (circles) or CA fibres (triangles). Note that slow decrease in apparent Congo red binding for CA-initiated CHIM polymerization mirrors CA self-seeded polymerization (inset). **b**, Effect of CHIM[SC] (circles) and CHIM[CA] (triangles) on CHIM polymerization. Representation of repeated seeding demonstrates the conformation propagation shown at top. **c**, Western blot analysis of first ([SC] and [CA]) and second ([[SC]] and [[CA]]) generation CHIM fibres incubated with (+) or without (-) chymotrypsin. Also shown is the digest of spontaneously polymerized CHIM fibres (SPONT). Relative molecular masses (M_r) are indicated.

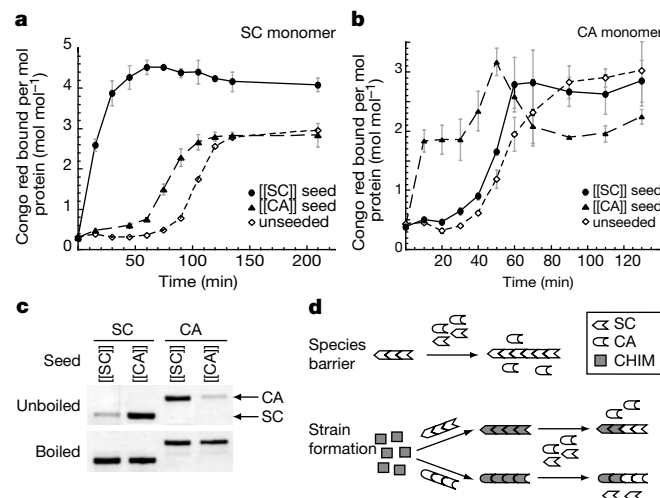


Figure 4 CHIM amyloids retain specificity of parent seed. **a**, **b** Second generation CHIM fibres ([[SC]] or [[CA]]) were used to seed either SC (**a**) or CA (**b**) polymerization. In the case of CA polymerization, a small glass bead was added to inhibit loss of apparent Congo red binding (see methods). **c**, SDS-PAGE analysis of polymerization reactions (SC or CA monomer), seeded with the indicated fibre. Aliquots were taken at 30 min, before spontaneous polymerization begins but after seeded polymerization was mostly completed. Appearance of band after boiling (lower) shows that loss of band in absence of boiling (upper) is due to polymerization, not loss of protein. **d**, Model depicting relationship between species barrier and prion conformation. Sequences with different strain conformation propensities (SC or CA) have a robust transmission barrier (top). A promiscuous prion (CHIM) can adopt either conformation (bottom), but once incorporated, retains the strain conformation and seeding specificity of the initial seed.

conformational differences between the fibres (Fig. 3c). Chymotrypsin patterns obtained from CHIM[CA] and CHIM[[CA]] were substantially different from those derived with CHIM[SC] and CHIM[[SC]]. These differences were retained throughout the time course of digestion (data not shown), arguing that these patterns reflected real conformational differences rather than differences in protease sensitivity. We conclude that distinct fibre conformations could be propagated with pure recombinant proteins.

We next examined the effect of these conformational differences on the seeding specificity of CHIM. We used second-generation fibres, which contain only 0.075% residual parent seed—an amount insufficient to affect conversion of either SC or CA polymerization (data not shown). As with spontaneously converted CHIM fibres, CHIM[[SC]] seeded SC monomer (Fig. 4a), whereas CHIM[[CA]] minimally affected the lag phase of SC polymerization. By contrast, CHIM[[CA]] gained the ability to seed CA monomer, whereas addition of either CHIM[[SC]] (Fig. 4b) or spontaneously polymerized CHIM (Fig. 2d) only modestly affected CA polymerization. We observed similar specificity when we monitored polymerization by thioflavin-T binding (data not shown) or formation of SDS-resistant aggregates (Fig. 4c). Therefore, the chimaera can adopt two distinct amyloid fibre conformations with different species preferences. The existence of two discernable fibre conformations suggests an explanation for why spontaneously formed CHIM fibres failed to seed CA *in vitro* (Fig. 2d): unseeded CHIM fibres have a preference for the SC-specific conformation. Consistent with this, the proteolytic pattern (Fig. 3c) and Congo red binding properties (Fig. 2a) of spontaneously polymerized CHIM resembled that of CHIM[SC] and CHIM[[SC]], arguing that they share a common conformation.

Originally attributed to differences in the nucleic acid sequence of a hypothetical slow virus, strains now appear to be an intrinsic feature of conformation-based inheritance in both mammalian^{1,2} and yeast prions^{3,16,17,23}. Two questions now arise. What are the molecular requirements for establishing and maintaining strains? What is the relationship between strains and cross-species transmission? Here we have established that a pure protein, devoid of covalent modifications⁷ and contaminants such as lipids or polysaccharides that might act as scaffolds^{27,28}, can adopt conformationally distinct fibres. These conformations are initially determined by the parent seed, but afterwards propagate stably. In the mammalian prion system strains are accompanied by differences in prion conformation^{1,5,8}, and these differences can be propagated in a partially purified *in vitro* system^{2,29}. Together, these observations strongly support the idea that a single protein can adopt several different self-propagating prion conformations and that these conformational differences are the basis for the diversity of strain phenotypes.

Our results also establish a link between a prion's conformation (and by inference the strain associated with it) and seeding specificity. The CHIM strain seeded by SC efficiently initiates polymerization of SC but has only modest effect on CA conversion, whereas CA-induced CHIM adopts a strain that resembles CA and specifically catalyses polymerization of CA. These data indicate that a substantial part of a robust species barrier, like that between *S. cerevisiae* and *C. albicans* (Fig. 4d), is due to the propensities of the two sequences for forming distinct non-interacting conformations. However, a promiscuous prion (in our case, CHIM) can access several conformations, but once templated faithfully propagates the specificity and conformation of the initial seed. Thus, species barriers and strains are intimately related phenomena: a prion's primary structure dictates the spectrum of favoured strain conformations, whereas a given prion strain infects only species capable of adopting prion conformations compatible with that strain.

The ability of strains to modulate prion specificity might help explain the apparent transmission of bovine spongiform

encephalopathy (BSE; or 'mad cow' disease) to many mammals, including human beings^{7,8}. On a more speculative note, it is possible that the process of infection, rendering and reinfection, which led to the BSE epidemic, also resulted in the selection and specific amplification of a prion strain with enhanced virulence³⁰. □

Methods

Plasmid construction

URA-marked inducer plasmids containing GFP fusions of SC, CA or CHIM (426CpSCGFP, 426CpCAGFP, 426CpSC(1–39)CAGFP respectively) were generated in a previous study¹⁷ using the copper-inducible CUP1 promoter for overexpression experiments. We generated integrative vectors by cloning the relevant prion-forming domains from these plasmids into URA-marked integration plasmids, together with the native promoter region of SUP35, a highly charged middle domain, and the C-terminal translation termination domain of native Sup35.

Gene integration and replacement

Integrative plasmids were linearized by restriction-enzyme digest (*MscI*) in the native promoter region, and transformed into *S. cerevisiae* yeast strain 74D-94 [*PSI*⁺] and [*psi*⁺]¹⁰. We replaced the native allele by gamma integration and used passage on medium containing 5-fluoro-orotic acid to select for recombination excision events. The presence of the expected prion-determining domain (either CA or CHIM) was detected by polymerase chain reaction (PCR), and western blot analysis confirmed proper expression of integrated domains.

Induction of prion states

Induction experiments were done as described¹⁷. Transformants with SC, CA or CHIM inducer plasmids were grown in SD-Ura medium and induced with 50 μ M CuSO₄ for 24 h. Saturated cultures were plated on SD-Ade medium and visible colonies were counted after 5 d growth at 30 °C. To determine strain phenotype, we grew inductants from SD-Ade plates overnight in liquid SD-Ade medium and plated them on non-selective low Ade medium. Colonies were counted and binned after 3 d growth.

Conversion to amyloid

For CA and CHIM a 9 $\times$ His-tagged protein, or for SC a 7 $\times$ His-tagged protein was purified in denaturing conditions as described¹⁷. Experiments with non-tagged protein showed similar results as with tagged protein (our unpublished data). Polymerization and Congo red binding assays have been described¹⁷. All seeding reactions used 3% mol mol⁻¹ of sonicated fibres. In CA experiments, with the exception of the inset of Fig. 3a, we added a small glass bead (3 mm in diameter), which inhibits loss of Congo red binding, perhaps by increasing shear, while only modestly altering polymerization kinetics. SDS solubility was assayed as described²⁴, with Coomassie staining for protein detection. Electron microscopy was done as before¹².

Limited proteolysis and western analysis

Fibres (10 μ M) in 150 mM NaCl, 5 mM KH₂PO₄, pH 7.4, were incubated with chymotrypsin (1/125 mol mol⁻¹) for 15 min at room temperature. Reactions were quenched by addition of SDS–PAGE loading buffer and incubation at 100 °C for 10 min. Digests were run on 16% Tris-Tricine gels (Novex) and transferred to nitrocellulose. Presence of protein was detected by immunoblotting with polyclonal antibodies to *S. cerevisiae* Sup35 prion-determining domain and detected by chemiluminescence (SuperSignal West Pico materials; Pierce).

Received 3 October; accepted 18 December 2000.

1. Prusiner, S. B. Prions. *Proc. Natl Acad. Sci. USA* **95**, 13363–13383 (1998).
2. Bessen, R. A. *et al.* Non-genetic propagation of strain-specific properties of scrapie prion protein. *Nature* **375**, 698–700 (1995).
3. Derkatch, I. L., Chernoff, Y. O., Kushnirov, V. V., Inge-Vechtomov, S. G. & Liebman, S. W. Genesis and variability of [PSI⁺] prion factors in *Saccharomyces cerevisiae*. *Genetics* **144**, 1375–1386 (1996).
4. Serio, T. R. & Lindquist, S. L. [PSI⁺]: an epigenetic modulator of translation termination efficiency. *Annu. Rev. Cell. Dev. Biol.* **15**, 661–703 (1999).
5. Telling, G. C. *et al.* Evidence for the conformation of the pathologic isoform of the prion protein enciphering and propagating prion diversity. *Science* **274**, 2079–2082 (1996).
6. Paushtkin, S. V., Kushnirov, V. V., Smirnov, V. N. & Ter-Avanesyan, M. D. Propagation of the yeast prion-like [psi⁺] determinant is mediated by oligomerization of the SUP35-encoded polypeptide chain release factor. *EMBO J.* **15**, 3127–3134 (1996).
7. Collinge, J., Sidle, K. C., Meads, J., Ironside, J. & Hill, A. F. Molecular analysis of prion strain variation and the aetiology of 'new variant' CJD. *Nature* **383**, 685–690 (1996).
8. Collinge, J. Variant Creutzfeldt-Jakob disease. *Lancet* **354**, 317–323 (1999).
9. Stansfield, I. *et al.* The products of the SUP45 (eRF1) and SUP35 genes interact to mediate translation termination in *Saccharomyces cerevisiae*. *EMBO J.* **14**, 4365–4373 (1995).
10. Chernoff, Y. O., Lindquist, S. L., Ono, B., Inge-Vechtomov, S. G. & Liebman, S. W. Role of the chaperone protein Hsp104 in propagation of the yeast prion-like factor [psi⁺]. *Science* **268**, 880–884 (1995).
11. Ter-Avanesyan, M. D., Dagkesamanskaya, A. R., Kushnirov, V. V. & Smirnov, V. N. The SUP35 omnipotent suppressor gene is involved in the maintenance of the non-Mendelian determinant [psi⁺] in the yeast *Saccharomyces cerevisiae*. *Genetics* **137**, 671–676 (1994).

12. DePace, A. H., Santoso, A., Hillner, P. & Weissman, J. S. A critical role for amino-terminal glutamine/asparagine repeats in the formation and propagation of a yeast prion. *Cell* **93**, 1241–1252 (1998).
13. Patino, M. M., Liu, J. J., Glover, J. R. & Lindquist, S. Support for the prion hypothesis for inheritance of a phenotypic trait in yeast. *Science* **273**, 622–626 (1996).
14. King, C. Y. *et al.* Prion-inducing domain 2–114 of yeast Sup35 protein transforms in vitro into amyloid-like filaments. *Proc. Natl Acad. Sci. USA* **94**, 6618–6622 (1997).
15. Glover, J. R. *et al.* Self-seeded fibers formed by Sup35, the protein determinant of [PSI], a heritable prion-like factor of *S. cerevisiae*. *Cell* **89**, 811–819 (1997).
16. Kushnirov, V. V., Kochneva-Pervukhova, N. V., Chechenova, M. B., Frolova, N. S. & Ter-Avanesyan, M. D. Prion properties of the Sup35 protein of yeast *Pichia methanolicus*. *EMBO J.* **19**, 324–331 (2000).
17. Santoso, A., Chien, P., Osherovich, L. Z. & Weissman, J. S. Molecular basis of a yeast prion species barrier. *Cell* **100**, 277–288 (2000).
18. Chernoff, Y. O. *et al.* Evolutionary conservation of prion-forming abilities of the yeast Sup35 protein. *Mol. Microbiol.* **35**, 865–876 (2000).
19. Liu, J. J. & Lindquist, S. Oligopeptide-repeat expansions modulate 'protein-only' inheritance in yeast. *Nature* **400**, 573–576 (1999).
20. Wickner, R. B. [URE3] as an altered URE2 protein: evidence for a prion analog in *Saccharomyces cerevisiae*. *Science* **264**, 566–569 (1994).
21. Sondheimer, N. & Lindquist, S. Rnq1: an epigenetic modifier of protein function in yeast. *Mol. Cell* **5**, 163–172 (2000).
22. Tuite, M. F., Mundy, C. R. & Cox, B. S. Agents that cause a high frequency of genetic change from [psi] to [psi⁺] in *Saccharomyces cerevisiae*. *Genetics* **98**, 691–711 (1981).
23. Zhou, P. *et al.* The yeast non-Mendelian factor [ETA] is a variant of [PSI], a prion-like form of release factor eRF3. *EMBO J.* **18**, 1182–1191 (1999).
24. Serio, T. R. *et al.* Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science* **289**, 1317–1321 (2000).
25. Paushkin, S. V., Kushnirov, V. V., Smirnov, V. N. & Ter-Avanesyan, M. D. In vitro propagation of the prion-like state of yeast Sup35 protein. *Science* **277**, 381–383 (1997).
26. Kochneva-Pervukhova, N. V. *et al.* Mechanism of inhibition of Psi prion determinant propagation by a mutation of the N-terminus of the yeast Sup35 protein. *EMBO J.* **17**, 5805–5810 (1998).
27. Appel, T. R., Dumpitak, C., Matthies, U. & Riesner, D. Prion rods contain an inert polysaccharide scaffold. *Biol. Chem.* **380**, 1295–1306 (1999).
28. Klein, T. R., Kirsch, D., Kaufmann, R. & Riesner, D. Prion rods contain small amounts of two host sphingolipids as revealed by thin-layer chromatography and mass spectrometry. *Biol. Chem.* **379**, 655–666 (1998).
29. Kocisko, D. A. *et al.* Species specificity in the cell-free conversion of prion protein to protease-resistant forms: a model for the scrapie species barrier. *Proc. Natl Acad. Sci. USA* **92**, 3923–3927 (1995).
30. Wilesmith, J. W., Ryan, J. B. & Atkinson, M. J. Bovine spongiform encephalopathy—Epidemiologic studies on the origin. *Vet. Rec.* **128**, 199–203 (1991).

Acknowledgements

We thank H. Bourne, H. Field, D. Julius, B. Panning, S. Lindquist, J. Reddy, S. Ribich, M. Scott and members of the Weissman and Lim lab for discussion and comments. This work was supported by the NIH, the Searle Scholars Program, the David and Lucile Packard Foundation, the Howard Hughes Medical Institute and a National Science Foundation Graduate Fellowship (P.C.).

Correspondence and requests for materials should be addressed to J.S.W. (e-mail: jsw1@itsa.ucsf.edu).

Increased dosage of a *sir-2* gene extends lifespan in *Caenorhabditis elegans*

Heidi A. Tissenbaum & Leonard Guarente

Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

In *Caenorhabditis elegans*, mutations that reduce the activity of an insulin-like receptor (*daf-2*)¹ or a phosphatidylinositol-3-OH kinase (*age-1*)² favour entry into the dauer state during larval development³ and extend lifespan in adults^{3–6}. Downregulation of this pathway activates a forkhead transcription factor (*daf-16*)^{7,8}, which may regulate targets that promote dauer formation in larvae and stress resistance and longevity in adults⁹. In yeast, the *SIR2* gene determines the lifespan of mother cells, and adding an extra copy of *SIR2* extends lifespan¹⁰. Sir2 mediates chromatin silencing through a histone deacetylase activity that depends on NAD (nicotinamide adenine dinucleotide) as a cofactor^{11–13}. We

have surveyed the lifespan of *C. elegans* strains containing duplications of chromosomal regions. Here we report that a duplication containing *sir-2.1*—the *C. elegans* gene most homologous to yeast *SIR2*—confers a lifespan that is extended by up to 50%. Genetic analysis indicates that the *sir-2.1* transgene functions upstream of *daf-16* in the insulin-like signalling pathway. Our findings suggest that Sir2 proteins may couple longevity to nutrient availability in many eukaryotic organisms.

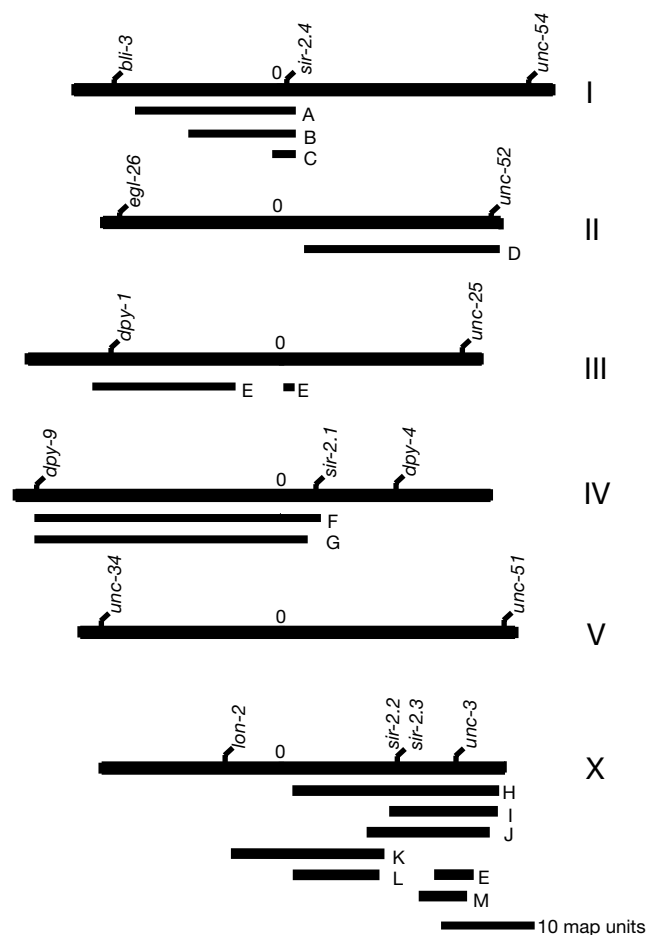


Figure 1 Genetic map of duplication strains. Genetic map showing some of the duplications tested for lifespan. Chromosome numbers are indicated to the right and drawn to scale according to WormBase (<http://www.wormbase.org/>) and Acedb (<http://elegans.swmed.edu/>). Chromosome sizes are as follows: I, 14,808,446 bp; II, 15,205,541 bp; III, 13,755,072 bp; IV, 17,452,072 bp; V, 21,150,598 bp; X, 17,727,995 bp. Thick line indicates chromosome; thin line represents duplication(s). In many cases, duplications were grouped together and assigned a letter A–M, as the endpoints were within one map unit. Mean lifespans of the duplication strains (in days) were tested in four trials. Trial 1: N2, 15.2 ± 0.5; KR1293-B, 11.7 ± 0.3; KR1725-C, 14.7 ± 0.3; KR1722-C, 15.8 ± 0.6; SP75-J, 14.0 ± 0.5; KR1704-C, 14.0 ± 0.4; TY1912-L, 15.8 ± 0.6; DR1786-F, 17.3 ± 1.0; SP116-I, 16.8 ± 0.5; SP117-H = 13.7 ± 0.6. Trial 2: N2, 15.1 ± 0.3; KR1732-C, 17.3 ± 0.5; RW6011-D, 17.6 ± 1.0; DR1786-F, 21.1 ± 1.0; TY1909-K, 14.0 ± 0.6; SP125-M, 18.8 ± 1.0. Trial 3: N2, 17.8 ± 0.6; MT7070-G, 14.6 ± 0.6; DR907-G, 16.6 ± 0.7; Trial 4: N2, 18.1 ± 0.8; KR1108-A, 18.7 ± 0.8; KR1110-A, 19.7 ± 0.7; KR1112-A, 18.7 ± 0.9; KR1236-C, 15.3 ± 0.6; KR1284-B, 12.8 ± 0.4; KR1280-B, = 10.2 ± 0.3; KR1548-C, 14.5 ± 0.4; KR1282-B, 10.9 ± 0.6; KR1815, 12.5 ± 1.0; SP1911-E, 14.3 ± 0.4 (see also Supplementary Information Table 1). *sir-2.2* and *sir-2.3* are adjacent to each other on the X chromosome. E is SP1911-*mnDp1*; this free duplication contains regions on both chromosome III (where it has a large deletion) and the X chromosome²⁷. The duplication in strain KR1815, *hDp30*, has not been precisely mapped. An additional nine strains that cover chromosome I and the X chromosome have been tested and do not display a significant increase in lifespan (not shown).

The *C. elegans* genome has four genes with similarity to yeast *SIR2* (ref. 14). Of these, we term the most related *sir-2.1* (31% identity to yeast *Sir2* in the conserved core domain). Other *sir-2* genes (*sir-2.2*, *sir-2.3* and *sir-2.4*) are between 10 and 20% identical in their conserved domains. To test whether these *sir-2* genes might regulate lifespan, we obtained strains with duplications of many chromosomal regions, including duplications covering these loci. Duplications are stably maintained in *C. elegans*, and the dosage of genes that they carry is increased by 50% (ref. 15). In total, we examined strains containing 35 different duplications, covering about half of the *C. elegans* genome (Fig. 1). Most of these strains, including those with extra copies of *sir-2.2*, *sir-2.3* and *sir-2.4*, exhibited lifespans similar to or shorter than the parental strain (see Supplementary Information Table 1). Notably, strain DR1786 containing the free duplication *mDp4* showed a significant extension in both mean and maximum lifespan (Fig. 2a; and Supplementary Information Table 1). This chromosome IV duplication includes the *sir-2.1* locus. Another chromosome IV duplication (*mDp1*; present in strains DR907 and MT7070) includes about 90% of the *mDp4* duplication, but is missing the region carrying the *sir-2.1* locus, and does not extend lifespan (Fig. 2a). Therefore, we inferred that *sir-2.1* or one of the other genes carried uniquely by *mDp4* extended the lifespan.

To test whether *sir-2.1* was responsible for this extension, we injected a 2.2-kilobase (kb) genomic fragment of *sir-2.1* into the N2 parental strain along with the transformation marker gene *rol-6* and identified three independent transgenic lines (see Methods). All three lines containing *sir-2.1* (*geEx1*, *geEx2* and *geEx3*) showed significantly extended lifespans ($P < 0.05$; mean lifespan 20.9 ± 0.3 , 27.4 ± 1.0 and 22.4 ± 0.4 days, respectively) compared with animals transgenic for *rol-6* (*su1066*) alone (mean lifespan 18.2 ± 0.3 days) or the N2 parent (mean lifespan 17.6 ± 0.2 days; Fig. 2b).

We derived stably integrated lines from animals carrying the *SIR2* extrachromosomal array (see Methods). Integrants derived from both *geEx1* (*geIn1*, *geIn2*) and *geEx2* (*geIn3*) had a several-fold increase in *sir-2.1* copy number, as determined by Southern blotting, compared with the N2 control (data not shown). These lines exhibited the same long lifespans as lines containing extrachromosomal arrays (*geEx1*, 20.9 ± 0.3 days; *geIn1*, 21.6 ± 0.3 ; *geIn2*, 21.5 ± 0.5 ; *geEx2*, 27.4 ± 1.0 ; *geIn3*, 27.6 ± 0.4 ; Fig. 2c). We conclude that, as in yeast, increasing the dosage of a *SIR2* gene extends the lifespan of *C. elegans*.

We considered whether *sir-2.1* extends lifespan by acting through one of the known *C. elegans* pathways. The extension in lifespan by mutations in the insulin-like signalling pathway (*daf-2*, *age-1* or *pdk-1*) is abolished in strains mutant in the downstream gene *daf-16* (refs 3, 6, 16); however, a second class of mutations that extend lifespan in *C. elegans*, termed *clk*, affects metabolic rate, characterized by a slowing in rhythmic behaviours and an increase in the time

Table 1 The *sir-2.1* transgene synergizes with TGF- β signalling mutants for dauer formation

Genotype	% dauer formation		Lifespan of array
	15°C (n)	20°C (n)	
<i>daf-4(m63)</i>	0.7 (449)	10.0 (981)	—
<i>daf-4(m63)</i> + pRF4	1.0 (209)	11.6 (353)	18.2
<i>daf-4(m63);geEx2</i>	15.0 (301)	48.0 (229)	27.4
<i>daf-4(m63);geEx3</i>	12.4 (431)	29.2 (380)	22.4
<i>daf-1(m40)</i>	0.6 (179)	2.7 (670)	—
<i>daf-1(m40)</i> + pRF4	0 (692)	0 (350)	18.2
<i>daf-1(m40);geEx2</i>	0 (30)	29.5 (285)	27.4
<i>daf-1(m40);geEx3</i>	1.3 (79)	13.0 (339)	22.4

Adult hermaphrodites were allowed to lay eggs for 4–18 h at 15°C after which they were removed and the plate was placed at either 15°C or 20°C. Eggs and L1 larvae were counted. Either 3 (20°C) or 6 (15°C) d later, plates were scored for dauer, nondauer, roller and nonroller. Data are the results of at least two completely independent trials. n, number of animals tested. As *daf-1* mutations are a maternal effect, hermaphrodites were allowed to lay eggs at 20°C for this mutant.

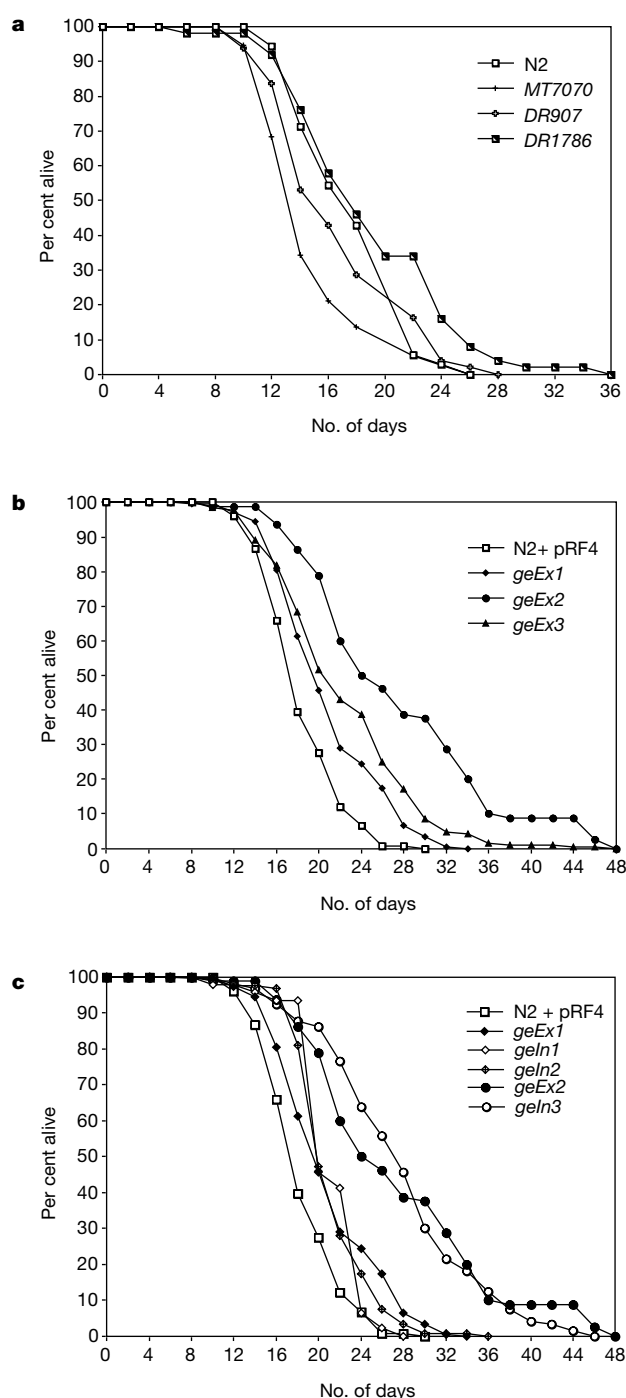


Figure 2 Increased dosage of *sir-2.1* extends lifespan. **a**, Lifespan curve shows that duplication of the *sir-2.1* region extends *C. elegans* lifespan. Data for each strain (mean $\pm$ s.e. (number of trials); n, number of animals scored): N2, 17.6 ± 0.2 (13), $n = 451$; MT7070, 14.6 ± 0.6 (1), $n = 38$; DR907, 16.6 ± 0.7 (1), $n = 49$; DR1786, 19.4 ± 0.6 (2), $n = 50$; $P < 0.05$ compared with N2. **b**, Lifespan curve shows introduction of *C. elegans* *sir-2.1* transgene extends lifespan. Data for each strain (mean $\pm$ s.e. (number of trials); n, number of animals scored): N2 + pRF4, 18.3 ± 0.3 (3), $n = 142$; *geEx1*, 20.9 ± 0.3 (6), $n = 214$; *geEx2*, 27.4 ± 1.0 (4), $n = 80$; *geEx3*, 22.4 ± 0.4 (6), $n = 209$. $P < 0.05$ for all three extrachromosomal array lines when compared with wild type. Also, there was no significant effect of *rol-6* on lifespan (mean lifespan 18.2 ± 0.3 with *rol-6*; 17.6 ± 0.2 wild-type alone). **c**, Integration of extrachromosomal array shows similar lifespan extension to the free array. Data for each strain (mean $\pm$ s.e. (number of trials); n, number of animals scored): N2 + pRF4, 18.3 ± 0.3 (3), $n = 142$; *geEx1*, 20.9 ± 0.3 (6), $n = 214$; *geIn1*, 21.6 ± 0.3 (2), $n = 46$; *geIn2*, 21.5 ± 0.5 (3), $n = 121$; *geEx2*, 27.4 ± 1.0 (4), $n = 80$; *geIn3*, 27.5 ± 0.4 (7), $n = 276$.

of development from embryo to adulthood. Mutations in this class are not suppressed by *daf-16* (ref. 17).

We thus tested whether *sir-2.1* may be acting in this pathway by determining the effect of a mutation in *daf-16* on *sir-2.1*-mediated longevity. The extrachromosomal array or integrated *sir-2.1* was crossed into a background containing *daf-16(mgDf50)*—a large deletion that eliminates most of the *daf-16* coding region⁷. *daf-16(mgDf50)* animals displayed a shorter mean lifespan than did wild-type animals (14.6 ± 0.1 days in *daf-16(mgDf50)* versus 17.6 ± 0.2 in N2) similar to other alleles of *daf-16* (refs 7, 8). The lifespan of *geln3* animals was 27.6 ± 0.4 days, and was reduced to 13.7 ± 0.2 in *daf-16(mgDf50);geln3* animals, similar to the lifespan of

daf-16(mgDf50) mutants alone (Fig. 3a). Similarly, when *daf-16(mgDf50)* was crossed to any of the three extrachromosomal arrays, lifespan was reduced to the level of *daf-16* mutant alone (Fig. 3a, legend). Therefore, the lifespan extension of the *sir-2.1* transgene requires *DAF-16* activity, suggesting that *sir-2.1* acts in the insulin-like signalling pathway.

If *sir-2.1* acts in this pathway, the transgene should not further extend the lifespan of a *daf-2* mutant, in which the pathway has been inactivated. *sir-2.1* transgenic arrays were thus crossed into a *daf-2(e1370)* mutant background. The *daf-2* mutation increased lifespan markedly, similar to previous data^{3,5,6,18} (Fig. 3b). However, no further extension was observed in a strain bearing the transgene and *daf-2(e1370)*; that is, the lifespan of the *daf-2(e1370);geln3* was similar to that of *daf-2(e1370)* animals (Fig. 3b). Thus, these experiments indicate that the *sir-2.1* transgene functions to extend lifespan in the insulin-like signalling pathway.

Many mutations in the insulin-like signalling pathway show other pleiotropies, including dauer formation, a reduced brood size and early larval lethality^{4,18–20}. Therefore, we determined whether *sir-2.1* extrachromosomal transgenic animals displayed any pleiotropic phenotypes. Brood size and time of development from hatching through the four larval stages to adulthood were normal in these animals when compared with wild-type animals carrying a *rol-6(su1066)* extrachromosomal array (see Supplementary Information Table 2). Furthermore, there was little dauer formation at 27 °C or at lower temperatures in any of the transgenic *SIR2* strains compared with wild-type animals with or without the *rol-6(su1066)* extrachromosomal array (data not shown). Thus, the *sir-2.1* transgene by itself does not affect fertility, early development or dauer formation.

Mutations in several genes in the insulin-like signalling pathway including *daf-2*, *unc-31* and *unc-64* also synergize with mutations in a parallel, transforming growth factor- β (TGF- β) signalling pathway, represented by *daf-4* and *daf-1* (refs 7, 20). *daf-1* and *daf-4* are type I and type II TGF- β receptors, and mutations in either of these genes cause a temperature-sensitive dauer-constitutive phenotype in larvae but do not affect lifespan in adults^{5,21}. Thus, to establish further that *sir-2.1* is linked to the insulin-like signalling pathway, we crossed the *sir-2.1* transgenic lines into *daf-4(m63)* or *daf-1(m40)* mutants.

We observed a striking increase in dauer formation at the permissive temperature of 20 °C in *sir-2.1* transgenic animals in combination with either *daf-4* or *daf-1* mutations (Table 1). Moreover, the tendency to form dauers correlated with the lifespan extension of the two transgenic lines tested. For example, at 20 °C 48% of the *daf-4;geEx2* animals formed dauers (*geEx2* increases lifespan 1.5 times), whereas 29% of the *daf-4;geEx3* animals formed dauers (*geEx3* increases lifespan 1.2 times). Even at 15 °C, both *daf-4;geEx2* and *daf-4;geEx3* animals showed a significant increase in dauer formation (Table 1). Thus, like mutations of other components of the insulin-like signalling pathway, the *sir-2.1* transgene synergizes with mutations in the TGF- β signalling pathway. As predicted, the *sir-2.1* transgene did not synergize with *daf-2(e1370)* for dauer formation (data not shown). These findings further bolster the claim that *sir-2.1* exerts its effect on the insulin-like signalling pathway. Evidently the effect of the *sir-2.1* transgene alone is too subtle to trigger dauer formation without the sensitizing *daf-1* or *daf-4* mutations.

In yeast, *Sir2* may act at the interface of nutrient availability and survival, by extending lifespan in response to calorie restriction²², and also by allowing haploid cells to mate and thus sporulate when starved. In *C. elegans*, *sir-2.1* may also couple environmental conditions to survival mechanisms; that is, the decision to entry into dauer in larvae, and the determination of lifespan in animals that proceed to adulthood. The mechanism by which *sir-2.1* functions may be the silencing of genes upstream of *daf-16* in cells that respond to the *DAF-2* ligand, the silencing of genes that result in

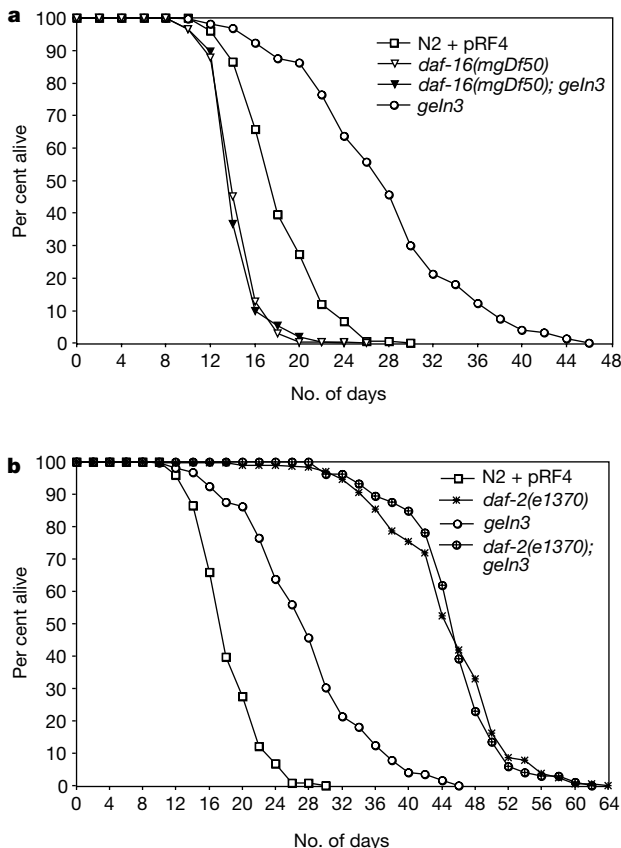


Figure 3 *sir-2.1* functions in the dauer signalling pathway. **a**, Lifespan curve shows that *daf-16* suppresses the lifespan extension of *sir-2.1* transgene. Data for each strain (mean $\pm$ s.e. (number of trials); *n*, number of animals scored): N2 + pRF4, 18.3 ± 0.3 (3), *n* = 142; *daf-16(mgDf50)*, 14.6 ± 0.1 (5), *n* = 274; *daf-16(mgDf50);geln3*, 13.7 ± 0.2 (4), *n* = 204; *geln3*, 27.5 ± 0.4 (7), *n* = 276. Similar findings were seen with the extrachromosomal array lines: *daf-16(mgDf50);geEx1*, 14.2 ± 0.5 (1), *n* = 40; *daf-16(mgDf50);geEx2*, 14.0 ± 0.3 (1), *n* = 29; *daf-16(mgDf50);geEx3*, 14.4 ± 0.4 (1), *n* = 47. This experiment was repeated on different independent isolates from both *daf-16(mgDf50);geEx1* and *daf-16(mgDf50);geEx3* with similar results (not shown). **b**, Lifespan curve shows that *daf-2* does not synergize with *sir-2.1* transgene for lifespan extension. Data for each strain (mean $\pm$ s.e. (number of trials); *n*, number of animals): N2 + pRF4, 18.3 ± 0.3 (3), *n* = 142; *daf-2(e1370)*, 44.7 ± 0.4 (6), *n* = 293; *daf-2(e1370);geln3*, 45.0 ± 0.5 (2), *n* = 105; *geln3*, 27.5 ± 0.4 (7), *n* = 276. Similar values were seen with another independent isolate of *daf-2(e1370);geln3* (not shown). When crossed into a *daf-2* background, *geEx1* and *geEx3* showed similar findings with the means similar to *daf-2(e1370)* alone: *daf-2(e1370);geEx1* = 41.9 ± 1.1 (2), *n* = 63; *daf-2(e1370);geEx3*, 47.2 ± 1.4 (1), *n* = 36. This experiment was repeated on different independent isolates from both *daf-2(e1370);geEx1* and *daf-2(e1370);geEx3* with similar results (not shown). Side by side comparisons showed that the *rol-6(su1066)* extrachromosomal array did not affect the mean or maximum lifespans of the *daf-2(e1370)* mutant (not shown).

production of the ligand in signalling cells, or both. Overexpression of *sir-2.1* would promote longevity and predispose animals to dauer formation by hyper-repressing these genes.

It seems remarkable that replicative ageing in yeast mother cells and postmitotic ageing in the soma of adult worms are both regulated by Sir2 proteins. In yeast, one beneficial effect of Sir2 appears to occur in the ribosomal DNA, where Sir2-silencing represses recombination and may also coordinate rRNA synthesis and growth rate. In *C. elegans*, silencing by Sir2 may couple nutrient availability to the level of signalling through the insulin-like signalling pathway. It will be of interest to determine whether Sir2 proteins regulate the rate of ageing in still higher eukaryotes, which contain both dividing and postmitotic cells in their soma. If so, SIR2 genes may provide a link that coordinates ageing in these two kinds of tissues to nutrient availability in an underlying and pervasive regulatory mechanism. □

Methods

Duplication strains

All strains were maintained and handled as described^{23,24}. Duplication strains were maintained by picking individual worms and examining the brood for proper segregants.

Transgenic worms

A 2.2-kb polymerase chain reaction (PCR) fragment containing the complete *sir-2.1* coding sequence and 400 base pairs (bp) upstream and 300 bp downstream was injected at 50 ng μl^{-1} along with pRF4 at 100 ng μl^{-1} (ref. 25) to obtain stable extrachromosomal transgenic lines. Lines were maintained by picking roller animals.

Integration

For each of *geEx1*, *geEx2* and *geEx3*, 10 L4 hermaphrodites were placed on 10 plates subjected to gamma irradiation from a Gammacell 220⁶⁰ Cosource at 4,000 rads. We then placed 5 animals onto each of 20 plates at 20° and allowed the plates to starve such that no more bacteria remained. Each of the 20 plates was then chunked onto a fresh plate. After 2 d, a total of 20 worms was singly picked to a plate from every parent plate for a total of 400 worms per gamma irradiation. After 4 d, plates were scored for the presence of rollers and non-rollers. We kept any plates with all rollers, and followed them for an integrated line. From these plates a total of five individual hermaphrodites were singled to new individual plates and scored 4 d later for all rollers. Integrants segregated all rollers on all of the five plates.

Strain construction

For *daf-2*, *daf-2(e1370)*, males were mated to the transgenic SIR2 lines at 15°C. After 5–7 d, we transferred putative roller cross progeny to individual plates at 25°C, and scored the plates 3 d later for the presence of dauers. Roller dauers were returned to 15°C to recover and singled to individual plates. For integrated lines, 20 animals were transferred from the brood of the recovered dauers to individual plates and their progeny scored for the presence of 100% rollers.

For *daf-16*, *daf-16(mgDf50)* males were mated to the *sir-2.1* lines at 15°C. After 5–7 d, putative roller cross progeny were singled to individual plates at 15°C. We singled 12 animals to plates from each of the F₁ plates and allowed them to lay eggs. After 4–5 d, PCR was done of the F₂ parent to determine the *daf-16* genotype using primers SO77 and SO78 (ref. 7). Strains were maintained by transferring roller animals to new plates.

For *daf-1/daf-4*, wild-type males were mated to either *daf-1(m40)* or *daf-4(m63)* hermaphrodites at 15°C. After 5–7 d, non-Daf F₁ male cross progeny were mated to the transgenic SIR2 lines at 15°C. After 5–7 d, 10 putative roller cross progeny hermaphrodites were singled to plates at 25°C. After 3 days, plates segregating roller dauer progeny were kept. Roller dauers were then picked off the plate and allowed to recover individually on a plate at 15°C to establish the strain. We maintained strains by transferring roller animals to new plates.

Lifespan and development assays

Lifespan assays were done at 20°C. Adult hermaphrodites were picked (4–10 per plate) from each strain and allowed to undergo one full generation at 15°C or 20°C. From these plates, we picked individual L4s or young adults to plates at 20°C containing 400 $\mu\text{g ml}^{-1}$ 5'-fluoro-2'-deoxyuridine (FUDR), which blocks DNA synthesis and causes animals to lay eggs that do not develop and eliminates the need to transfer animals throughout the lifespan assay²⁶. Animals were tapped every 2–4 d and were scored as dead when they did not move after repeated taps with a pick. A limited number of experiments were carried out on plates without FUDR by transferring adult animals every 1–2 d to new plates and these revealed the same, long lifespans of *sir-2.1* transgenic animals. All statistical tests were done using JMP 4.0 software.

Timing of development was scored for each strain at 20°C. Plates were scored every 12–18 h for developmental stage. No noticeable differences were observed comparing the roller strains with and without the *sir-2.1* transgene. For 27°C dauer formation, animals

were allowed to lay eggs at 27°C for 4 h. We scored plates 2 d later for the presence of dauers and non-dauers. Both wild-type and *unc-31(e928)* animals were used as controls for dauer induction.

Received 1 September 2000; accepted 2 January 2001.

- Kimura, K., Tissenbaum, H. A., Liu, Y. & Ruvkun, G. The *daf-2* insulin receptor family member regulates longevity and diapause in *Caenorhabditis elegans*. *Science* **277**, 942–946 (1997).
- Morris, J. Z., Tissenbaum, H. A. & Ruvkun, G. A phosphatidylinositol-3-OH kinase family member regulating longevity and diapause in *Caenorhabditis elegans*. *Nature* **382**, 536–539 (1996).
- Larsen, P. L., Albert, P. S. & Riddle, D. L. Genes that regulate both development and longevity in *Caenorhabditis elegans*. *Genetics* **139**, 1567–1583 (1995).
- Friedman, D. B. & Johnson, T. E. A mutation in the *age-1* gene in *Caenorhabditis elegans* lengthens life and reduces hermaphrodite fertility. *Genetics* **118**, 75–86 (1988).
- Kenyon, C., Chang, J., Gensch, E., Rudner, A. & Tabtiang, R. A *C. elegans* mutant that lives twice as long as wild type. *Nature* **366**, 461–464 (1993).
- Dorman, J. B., Albinder, B., Shroyer, T. & Kenyon, C. The *age-1* and *daf-2* genes function in a common pathway to control the lifespan of *Caenorhabditis elegans*. *Genetics* **141**, 1399–1406 (1995).
- Ogg, S. *et al.* The Fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in *C. elegans*. *Nature* **389**, 994–999 (1997).
- Lin, K., Dorman, J. B., Rodan, A. & Kenyon, C. *daf-16*: An HNF-3/forkhead family member that can function to double the life-span of *Caenorhabditis elegans*. *Science* **278**, 1319–1322 (1997).
- Lithgow, G. J., White, T. M., Melov, S. & Johnson, T. E. Thermotolerance and extended life-span conferred by single-gene mutations and induced by thermal stress. *Proc. Natl Acad. Sci. USA* **92**, 7540–7544 (1995).
- Kaeberlein, M., McVey, M. & Guarente, L. The SIR2/3/4 complex and SIR2 alone promote longevity in *Saccharomyces cerevisiae* by two different mechanisms. *Genes Dev.* **13**, 2570–2580 (1999).
- Imai, S., Armstrong, C. M., Kaeberlein, M. & Guarente, L. Transcriptional silencing and longevity protein Sir2 is an NAD-dependent histone deacetylase. *Nature* **403**, 795–800 (2000).
- Smith, J. S. *et al.* A phylogenetically conserved NAD⁺-dependent protein deacetylase activity in the Sir2 protein family. *Proc. Natl Acad. Sci. USA* **97**, 6658–6663 (2000).
- Landry, J. *et al.* The silencing protein SIR2 and its homologs are NAD-dependent protein deacetylases. *Proc. Natl Acad. Sci. USA* **97**, 5807–5811 (2000).
- Frye, R. A. Phylogenetic classification of prokaryotic and eukaryotic Sir2-like proteins. *Biochem. Biophys. Res. Comm.* **273**, 793–798 (2000).
- Herman, R. K., Madl, J. E. & Cari, C. K. Duplications in *Caenorhabditis elegans*. *Dev. Biol.* **49**, 200–219 (1979).
- Paradis, S., Ailion, M., Tokar, A., Thomas, J. H. & Ruvkun, G. A PDK1 homolog is necessary and sufficient to transduce AGE-1 PI3 kinase signals that regulate diapause in *Caenorhabditis elegans*. *Genes Dev.* **13**, 1438–1452 (1999).
- Lakowski, B. & Hekimi, S. Determination of life-span in *Caenorhabditis elegans* by four clock genes. *Science* **272**, 1010–1013 (1996).
- Gems, D. *et al.* Two pleiotropic classes of *daf-2* mutation affect larval arrest, adult behavior, reproduction and longevity in *Caenorhabditis elegans*. *Genetics* **150**, 129–155 (1998).
- Tissenbaum, H. A. & Ruvkun, G. An insulin-like signaling pathway affects both longevity and reproduction in *Caenorhabditis elegans*. *Genetics* **148**, 703–717 (1998).
- Ailion, M., Inoue, T., Weaver, C. L., Holdcraft, R. W. & Thomas, J. H. Neurosecretory control of aging in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **96**, 7394–7397 (1999).
- Riddle, D. L. & Albert, P. S. in *C. elegans II* (eds Riddle, D. L., Blumenthal, T., Meyer, B. J. & Priess, J. R.) 739–768 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1997).
- Lin, S. J., Defossez, P. A. & Guarente, L. Lifespan extension in *S. cerevisiae* by calorie restriction requires NAD and SIR2. *Science* **289**, 2126–2128 (2000).
- Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
- Sulston, J. & Hodgkin, J. in *The Nematode Caenorhabditis elegans* (ed. Wood, W. B.) 587–602 (Cold Spring Harbor Laboratory, Cold Spring Harbor, 1988).
- Mello, C. C., Kramer, J. M., Stinchcomb, D. & Ambros, V. Efficient gene transfer in *C. elegans*: extrachromosomal maintenance and integration of transforming sequences. *EMBO J.* **10**, 3959–3970 (1991).
- Apfeld, J. & Kenyon, C. Cell nonautonomy of *C. elegans daf-2* function in the regulation of diapause and lifespan. *Cell* **95**, 199–210 (1998).
- Hedgecock, E. M. & Herman, R. K. The *ncl-1* gene and genetic mosaics of *Caenorhabditis elegans*. *Genetics* **141**, 989–1006 (1995).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank P. Garrity, M. Hunter-Ensor, C. Ceol, O. Hobert, F. Slack, G. Ruvkun, M. Kaeberlein, M. McVey, M. Alkema and A. Dines for suggestions, encouragement and discussions throughout this work; E. Ford, M. Alkema, B. Hersh and M. Hunter-Ensor for critically reading the manuscript; C. Wolkow for the FUDR protocol; O. Hobert for the integration protocol; S. Moseley for help with statistical analysis; H. R. Horvitz for providing lab space where this work was done, and N. An and H. R. Horvitz for providing some of the strains used in this study; G. Ruvkun and P. Delorme for use and help with the microinjection scope; and G. Ruvkun and S. Kennedy for the primer sequence of SO77 and SO78. Many strains were sent by T. Stiernagle at the *Caenorhabditis* Genetics Center which is funded by the National Institutes of Health National Center for Research Resources. H.A.T. is funded by the Helen Hay Whitney Foundation. L.G. is funded by grants from the NIH, the Ellison Medical Foundation, and the Howard and Linda Stern Fund.

Correspondence and requests for materials should be addressed to L.G. (e-mail: leng@MIT.edu).

GTPase activity of dynamin and resulting conformation change are essential for endocytosis

Bruno Marks^{*†}, Michael H. B. Stowell^{*†}, Yvonne Vallis^{*†}, Ian G. Mills^{*}, Adele Gibson[‡], Colin R. Hopkins[‡] & Harvey T. McMahon^{*}

^{*} MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

[‡] MRC Laboratory for Molecular Cell Biology, University College London, Gower Street, London WC1E 6BT, UK

[†] These authors contributed equally to this work

Dynamin is a large GTPase with a relative molecular mass of 96,000 (M_r 96K) that is involved in clathrin-mediated endocytosis and other vesicular trafficking processes^{1,2}. Although its function is apparently essential for scission of newly formed vesicles from the plasma membrane, the nature of dynamin's role in the scission process is still unclear^{3,4}. It has been proposed that dynamin is a regulator (similar to classical G proteins) of downstream effectors⁵. Here we report the analysis of several point mutants of dynamin's GTPase effector (GED) and GTPase domains. We show that oligomerization and GTP binding alone, by dynamin, are not sufficient for endocytosis *in vivo*. Rather, efficient GTP hydrolysis and an associated conformational change are also required. These data argue that dynamin has a mechanochemical function in vesicle scission.

Temperature-sensitive mutants of the *Drosophila* homologue of dynamin, *shibire*, have a paralytic phenotype at the non-permissive temperature due to a reversible blockade of synaptic vesicle recycling^{6,7}. An accumulation of deeply invaginated membrane profiles, apparently of vesicles unable to detach from the plasma membrane, is observed in the paralysed flies⁸. On the basis of the *shibire* phenotype it was proposed that dynamin is involved in scission of nascent endocytic vesicles from the plasma membrane. This hypothesis has been strengthened by overexpression studies using dynamin mutants^{9–11}, and by the observation of a *shibire*-like blockade of endocytosis in GTP- γ S-treated synaptosomes¹².

In vitro, dynamin tubulates acidic phospholipid vesicles by helical oligomerization^{13,14}. Notably, the GTPase activity of dynamin is increased up to 1,000-fold by its adoption of such a helical conformation¹⁵. On these lipid nanotubes, GTP hydrolysis by dynamin results in a change in its conformation that can be observed as a dramatic increase in helical pitch (from 11 ± 2 to 20 ± 3 nm)¹⁵. Some studies have also reported fission of lipid tubules by dynamin (possibly involving a pinching/constriction action)^{13,16}. These observations and others suggest that dynamin is a mechanochemical enzyme that uses energy from GTP hydrolysis to generate the force required for membrane scission. However, recent work on GTPase-effector-domain (GED, see Fig. 1) mutants has cast doubt on this hypothesis by demonstrating that dynamin's GTPase activity (stimulated by oligomerization) is not required for endocytosis⁵. This has led to the alternative hypothesis that dynamin is a regulator of downstream effectors of scission, rather than being directly responsible for it. In an effort to distinguish between these two hypotheses, we designed mutants of dynamin that would enable us to test the relationship between dynamin's GTPase activity and its ability to support endocytosis.

Previous studies have implicated the GED of dynamin (Fig. 1) as the physiological activator of dynamin's GTPase activity^{5,17}. On the basis of the assumption that the dynamin GED is a GTPase-activating protein (GAP), we mutated all arginine and lysine residues in the GED. Mutants were tested using an *in vivo* assay for endocytosis (see Methods). Within the sensitivity range of our assay, none of the GED mutants seemed to have a significant

inhibitory effect on endocytosis of transferrin (Fig. 2; and data not shown). We chose to examine the R725A and K694A mutants of the GED further, as these are the GED mutants known to inhibit dynamin's GTPase activity (stimulated by oligomerization)⁵. We found that the GTPase activity of these mutants was comparable to that of wild-type dynamin (Fig. 2 and Table 1). In our GTPase assay we used lipid nanotubes as the template for oligomerization-dependent stimulation of dynamin's GTPase activity. Previously, the activity of K694A and R725A were assayed using microtubules as the template for oligomerization⁵. We have also assayed these mutants on microtubules and show that K_{cat} (catalytic constant) values for wild type, K694A and R725A are not significantly different (see Supplementary Information). We therefore conclude that dynamin's GED is probably not a GAP in the classical sense (that is, contributing a catalytic residue to the active site). More data and further discussion of these mutants can be found in the Supplementary Information. Notably, several well characterized motor proteins, such as kinesin and myosin, do not use an arginine finger in their catalytic mechanism. The structure of a large dynamin-like GTPase GTP-binding protein has been determined¹⁸, which showed that the phosphate-binding site was completely shielded from solvent, such that a potential GTPase-activating protein cannot approach. Thus, it is not unexpected that we cannot identify a catalytic arginine or lysine residue in dynamin's GED.

What is the possible function of dynamin's GED? In limited proteolysis experiments, the GED of dynamin co-purifies with the GTPase domain, and is probably involved in dynamin oligomerization^{5,17}. Deletion analysis of the GED has strengthened this hypothesis¹⁹. Furthermore, mutations in the GED reverse the reticular phenotype (oligomerized state) of a mutant (T65A; see below) (Fig. 2), supporting the proposed function of this domain in oligomerization.

Guided by the structure of Ras GTPase²⁰, we made several point mutations in and around the GTP-binding motifs (Fig. 1),

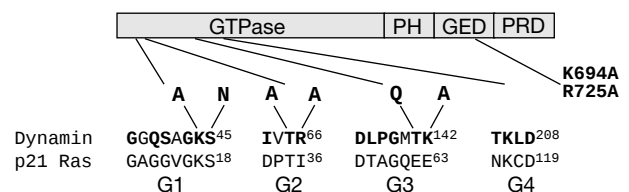


Figure 1 Domain structure of dynamin showing mutants of the GTPase domain. The nucleotide-binding sequences (G1–G4) of dynamin and p21 Ras are compared. Point mutations that we made are shown above the dynamin sequence. The bold residues in dynamin are conserved in dynamin family proteins from *Arabidopsis* to mammals. In addition to a GTPase domain, dynamins possess three other distinct and conserved domains. The pleckstrin homology (PH) domain binds to PtdIns(4,5)P₂-containing lipids^{24,25}. The GTPase effector domain (GED) is proposed to be involved in dynamin oligomerization and also activation of dynamin's GTPase activity^{5,17,19,26}. The proline/arginine-rich domain (PRD) at the carboxy terminus binds to many SH3-domain-containing proteins including amphiphysin and endophilin^{27–29}.

Table 1 Kinetic parameters K_{cat} and K_m for a number of dynamin mutants

Dynamin identity	Turnover rate (K_{cat}) (s)	K_m (μ M)
Wild type	3.1 ± 0.5	7.8 ± 2.5
S45N	0.083 ± 0.043	ND
T65A	0.023 ± 0.015	9.1 ± 1.1
R66A	0.45 ± 0.19	65.0 ± 5.3
T141Q	0.55 ± 0.15	7.6 ± 3
K142A	1.8 ± 0.4	6.7 ± 1.5
K694A	2.7 ± 0.5	ND
R725A	2.8 ± 0.2	4.3 ± 1.8

As the absolute specific activity (K_{cat}) (mean $\pm$ s.d.) varied with different preparations of dynamin and lipid nanotubes, the activities of the mutants were calculated as a percentage of wild-type activity determined in the same experiment (see Fig. 2). The absolute activities have been calculated using those percentages and a value for wild-type activity averaged over several different dynamin and lipid nanotube preparations. ND, not determined.

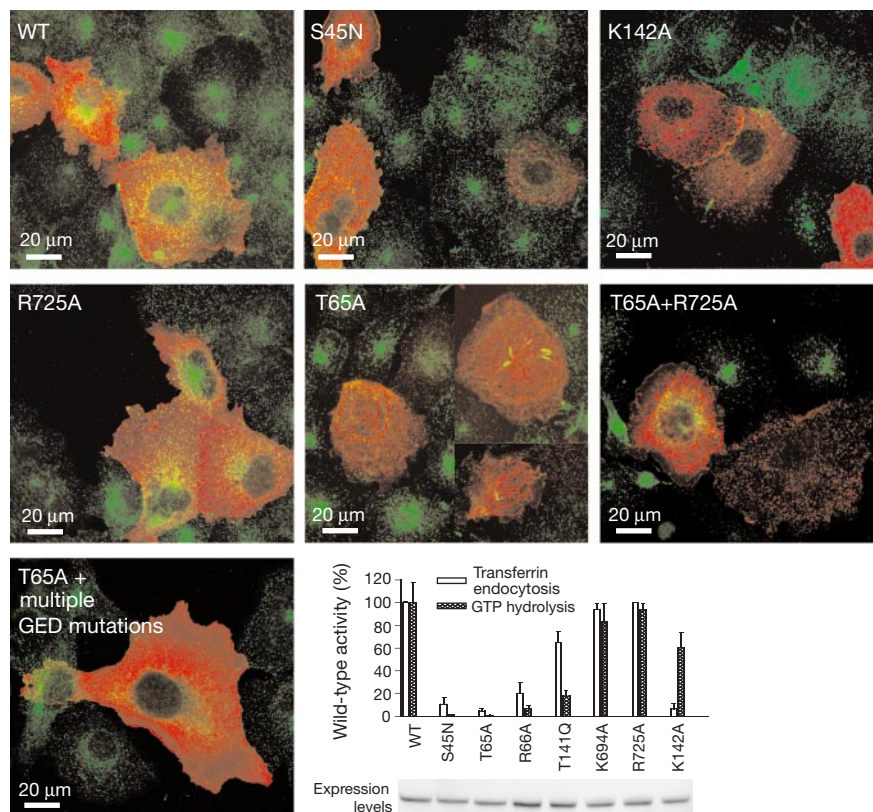


Figure 2 Uptake of biotinylated transferrin (green staining) into COS-7 fibroblasts, transiently overexpressing dynamin or dynamin mutants (red staining). The percentage of transfected cells showing wild-type (WT) transferrin uptake was quantified from several transfection experiments, and is shown in the graph along with a blot showing that there is similar expression of all the dynamin mutants (see Methods). With inhibitory mutants, even weakly expressing cells are strongly inhibited for transferrin uptake (see examples for S45N and K142A). Filled bars show GTP hydrolysis for the purified mutant proteins in the presence of 75 μ M GTP. We have presented a number of T65A-expressing cells

showing a reticular distribution of dynamin and worms of internalized transferrin. This distribution of T65A dynamin is progressively abolished by mutations of the oligomerization domain. Thus, T65A + R725A show no evidence for worms of transferrin and a reduced tubular appearance. T65A + F698A + R724E + R725E (multiple GED mutations) shows a predominantly cytoplasmic distribution of dynamin similar to WT dynamin transfections. This mutant also shows some uptake of transferrin at a perinuclear localization.

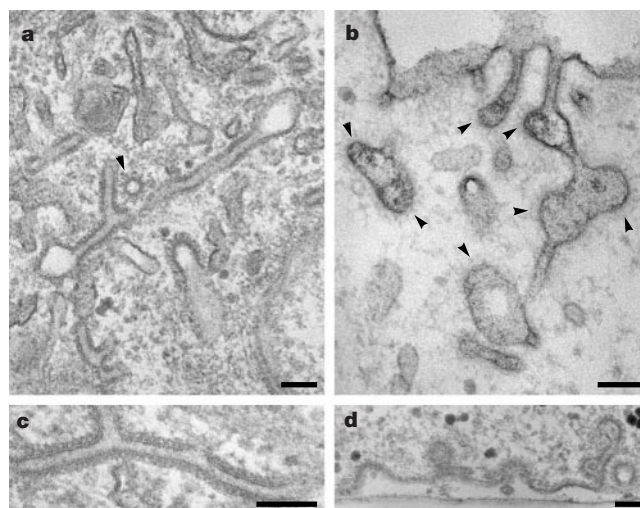


Figure 3 Thin-section electron microscopy of T65A-transfected COS cell. **a**, T65A-transfected cells are full of membrane tubules, many of which emanate from the plasma membrane. Frequently tubules are clearly surrounded by electron density, reminiscent of that seen with dynamin in GTP- γ S-treated synaptosomes or on lipid tubules *in vitro*. Cross-sections of these tubules show roughly 10 repeating 'spokes' (see arrowhead). **b**, Plasma-membrane invaginations at the cell surface of cells expressing T65A. Cells were stained with ruthenium red, outlining the continuities with the plasma membrane. Some domains (arrowheads) are clathrin coated. **c**, Tubulated membrane surrounded by

electron density that we propose to be dynamin (as no other proteins are being overexpressed in these cells). This correlates with the filamentous dynamin staining seen by immunofluorescence. This represents strong evidence that dynamin can tubulate the plasma membrane *in vivo* as has been shown *in vitro*¹⁴. **d**, At the plasma membrane of T65A-transfected cells, various intermediate stages of endocytosis were 'captured'. This is reminiscent of what is seen in *shibire Drosophila* at the non-permissive temperature. Scale bar, 0.1 μ m.

concentrating on residues that are thought to be directly or indirectly (through coordination of water) involved in catalysis. Initially, mutants were screened using our *in vivo* endocytosis assay. Several inhibitory mutants were identified (Fig. 2). These mutants showed a corresponding inhibition when assayed for GTPase activity, with the exception of K142A (Fig. 2 and Table 1). Further analysis showed that the S45N mutant is defective in GTP binding, as has been predicted^{10,11} (see Supplementary Information; and data not shown). Kinetic analysis revealed that R66A has a K_m (Michaelis constant) significantly higher than that of wild-type dynamin (Table 1). The K_m of the other mutants, K142A, T65A and T141Q, are similar to that of wild-type dynamin (Table 1), and are well below the intracellular GTP concentration of 50–150 μM (ref. 21).

If dynamin is a mechanochemical enzyme, GTP hydrolysis must be essential for its function. However, if dynamin acts as a molecular switch in a manner similar to that of classical G proteins, GTP–dynamin would represent the ‘on’ state, allowing endocytosis to proceed, and GTP hydrolysis would serve to switch it ‘off’. A mutant that could bind but not effectively hydrolyse GTP should enable one to distinguish between these two possibilities. The T65A mutant has a very low GTPase activity, but has a K_m close to that of the wild-type dynamin. The fact that this mutant strongly inhibits endocytosis implies that effective GTP hydrolysis by dynamin, and not simply GTP binding, is necessary for vesicle scission.

Of note, in our *in vivo* endocytosis assay, roughly 50% of cells expressing T65A had a very distinctive appearance (seen also to a lesser extent in T141Q and R66A, see Supplementary Information).

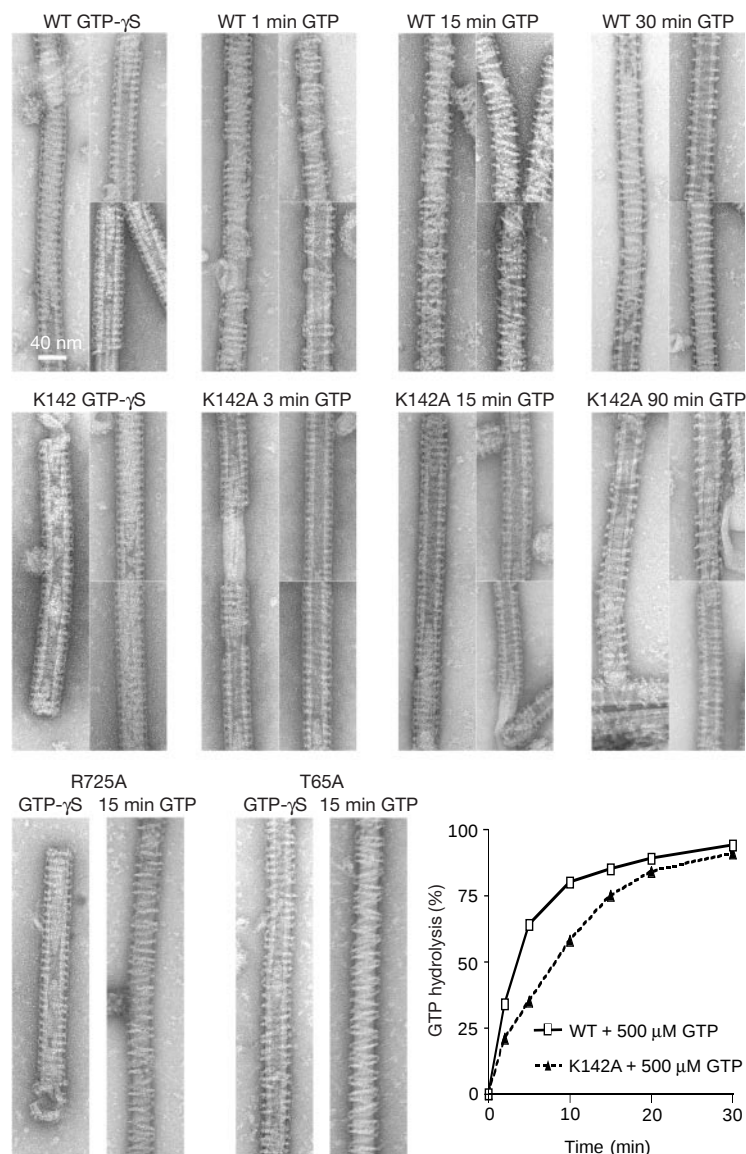


Figure 4 Conformational changes in dynamin on GTP hydrolysis. Purified, expressed dynamin bound to lipid nanotubes were visualized by negative staining in the presence of 100 μM GTP- γS or during the hydrolysis of 500 μM GTP. To compare the extent of GTP hydrolysis at different sampling times for WT versus K142A dynamin, aliquots of the electron microscopy (e-m) reactions were taken for TLC in addition to being put on e-m grids (graph, bottom right). The R725A mutant has essentially the same time-course of GTP hydrolysis as WT dynamin (see Table 1 and Supplementary Information). WT, K142A and R725A dynamin all adopt an ordered, tight, helical conformation in the presence of GTP- γS . In the presence of GTP, much of this order is immediately lost by WT (or R725A)

dynamin even at very short time points after hydrolysis. At long time points, WT dynamin is bound in a loose helix about the lipid nanotubes. In contrast, the K142A mutant stays in a conformation similar to that seen with GTP- γS , until most of the GTP has been hydrolysed. T65A dynamin does not adopt a tight, ordered conformation in the presence of GTP- γS , it is looser and disordered. This conformation does not apparently change as a result of GTP hydrolysis. The conformational changes reported here for baculovirus-expressed WT bovine dynamin I are identical to those reported in our previous paper for purified rat brain dynamin, which contains an equal mixture of 96K and 93K spliceforms.

The dynamin in these T65A-expressing cells had the appearance of being localized along thin filaments running throughout the cell (reticular, rather than the predominant cytosolic distribution seen with wild-type protein) (Fig. 2). Although there was no general transferrin staining within T65A-transfected cells (indicating an inhibition of endocytosis) many cells showed discreet, brightly stained 'worms' of transferrin (Fig. 2, T65A panel). Using thin-section electron microscopy, we have shown that the T65A-transfected cells are full of membrane tubules, many of which can be seen to be connected with the plasma membrane (Fig. 3). The tubules are of uniform diameter and are coated with rings of protein that have the same T-bar-like structures as are seen with dynamin bound to lipid nanotubes *in vitro* (compare Fig. 3c to purified dynamin on tubules in Fig. 4). We propose that the T65A dynamin oligomerizes as normal about the necks of invaginating vesicles in its GTP-bound state, but cannot hydrolyse GTP effectively, and therefore does not catalyse the scission reaction. Furthermore, as GTP hydrolysis is so impaired, the T65A dynamin has a much longer lifetime on the plasma membrane than would wild-type dynamin, and is therefore able to form long oligomers that create tubular membrane invaginations. We believe that the worms of transferrin that are sometimes observed in T65A-expressing cells are transferrin that has undergone endocytosis at the end of these tubular invaginations. Some of these invaginations may fuse (as endocytic vesicles would to form early endosomes) resulting in a transferrin-rich compartment that is still contiguous with the extracellular environment. We have used ruthenium red staining to confirm the surface accessibility of much of the tubulated membrane (Fig. 3). Using immuno-gold labelling, we observed that transferrin is often concentrated at the intracellular ends of the tubules (data not shown).

The T65A mutant supports the hypothesis that further to formation of a dynamin-GTP oligomer around a vesicle neck, GTP hydrolysis is necessary for scission to occur. Another mutant that we identified, T141Q, further supports this hypothesis. T141Q is partially inhibited in its GTPase activity and correspondingly it partially inhibits endocytosis (Fig. 2 and Table 1). Owing to the limitations of our endocytosis assay and the conservative criteria by which inhibition is 'scored', the extent of partial inhibition by T141Q is probably underestimated (see Methods). Nevertheless, for these mutants, there is a direct correlation between the degree of endocytic inhibition and inhibition of their GTPase activity, an expected result if GTP hydrolysis is directly required in the process of vesicle scission.

In contrast, the K142A mutant has only a moderately impaired GTPase activity (Table 1), but inhibits endocytosis strongly (Fig. 2). Initially, this finding seems to be incompatible with currently proposed hypotheses for dynamin's action. Because of this apparent anomaly, we chose K142A for further investigation. Previously we have shown that wild-type dynamin, oligomerized into a helix about PIP₂-containing lipid nanotubes, extends its pitch on GTP hydrolysis¹⁵. Figure 4 shows that with wild-type dynamin, a conformational change resulting in an average increase in helical pitch from 11 ± 2 nm to 15 ± 2 nm (mean $\pm$ s.d.) is evident at the shortest GTPase assaying times that we tested (after 1 min, which equates to about 25% GTP hydrolysis). With K142A, after equivalent GTP hydrolysis (3 min or 25% hydrolysis), the pitch of the dynamin helix had not changed significantly from that observed with K142A-GTP- γ S (11 ± 2 nm and 11 ± 1 nm, respectively). In fact, K142A dynamin did not change conformation from a tightly wound helical state until all of the GTP had been hydrolysed (which will never occur *in vivo*). This suggests that the catalytic cycle of K142A is defective in such a way that if there is any GTP available to bind, this occurs without the previous transition through a loosely helical GDP state. At present we do not have a precise understanding of the mechanistic defect present in K142A, but these observations suggest that as well as efficient GTP hydrolysis, a conformational change in dynamin is required for endocytosis, that is, mechano-

chemical coupling. The conformational change could be coupled to one of several steps in the catalytic cycle, such as hydrolysis, product release and/or a subsequent isomerization step.

We also examined the conformations of other dynamin mutants bound to lipid nanotubes. R725A behaved exactly as wild-type dynamin (Fig. 4). K694A, although not examined as thoroughly, also showed wild-type characteristics (data not shown). The T65A mutant did not form a tight, ordered conformation with GTP- γ S (average pitch 12 ± 2 nm). Its inability to hydrolyse GTP rapidly could be related to this.

It is interesting to note that after extensive GTP hydrolysis by wild-type dynamin we observe an increase in the percentage of dynamin showing increased helical pitch; however, we rarely observe the uniformity of conformation seen when dynamin is incubated with GDP¹⁵. This may be because GTP is never completely depleted in our experiments.

We have demonstrated a positive correlation between the GTPase activity of dynamin and its ability to support endocytosis *in vivo*. In addition, an anomaly to this correlation, K142A, was found to be defective in its ability to change conformation while still maintaining high GTPase activity. Thus, dynamin's function in endocytosis requires both GTP hydrolysis and a resulting conformational change before, or concomitant with, vesicle scission. $\square$

Methods

Transferrin uptake

We transfected COS-7 fibroblasts using DEAE dextran. Thirty-six hours after transfection, cells were transferred to serum-free medium and incubated overnight. The next day, the cells were incubated for 30 min at 37 °C in the presence of 25 μ g ml⁻¹ biotinylated human transferrin (Sigma), and then fixed (4% paraformaldehyde). Transferrin uptake was visualized using FITC-conjugated streptavidin (green) on an MRC 1024 scanning confocal microscope. We detected dynamin using an anti-dynamin polyclonal antibody (Ra19), followed by Texas Red-conjugated anti-rabbit antibody (Chemicon). Inhibition of endocytosis in transfected cells was defined as transferrin uptake less than 20% of that seen in untransfected cells in the same field. This ensures that only strongly inhibited cells are scored. A consequence of this is that the degree of inhibition of partially inhibitory mutants might be underestimated (for example, T141Q). In all mutants that were inhibited in transferrin uptake, dynamin no longer had a predominantly cytoplasmic distribution. With cells expressing T65A mutant dynamin, a filamentous appearance could be observed. The density and length of filaments was expression-level dependent. In Fig. 2, T65A is a serial reconstruction, rather than a single confocal section, to better illustrate these observations.

Dynamin expression and purification

Wild-type dynamin (bovine dynamin 1) and mutants were expressed using the baculovirus pBac 4 system (Novagen) and purified using a recombinant, bacterially expressed, GST-tagged, amphiphysin-2 SH3 domain as an affinity ligand, as described¹⁵. The quality of dynamin was regularly checked by SDS-PAGE and Coomassie staining, and dynamin was not used if more than roughly 5% degraded.

GTPase assays

GTPase assays were carried out both as described, using thin-layer chromatography (TLC)¹⁵, and by a continuous enzyme-coupled method for monitoring phosphate release (EnzChek, Molecular Probes). With both methods reactions were as follows. Reaction volumes were 20 μ l for the TLC assay and 60 μ l for the enzyme-coupled assay. Lipid nanotubes (0.1 mM lipid) and dynamin (typically 0.5 μ M) were incubated for 10 min at room temperature in buffer A (135 mM NaCl, 5 mM KCl, 20 mM HEPES, 1 mM MgCl₂) and then reactions were initiated by the addition of GTP (5–500 μ M). For the continuous enzyme-coupled assay, purine nucleotide phosphorylase (20 U ml⁻¹) and 2-amino-6-mercapto-7-methylpurine ribonucleoside (MESG) (200 μ M) were also present in the reaction mix from the beginning. After initiation of the reaction its progress was monitored as an increase in A₃₅₅ (due to the phosphate-dependent conversion of MESG to ribose-1-phosphate and 2-amino-6-mercapto-7-methylpurine) using a Hewlett Packard spectrophotometer. For the TLC assay, the GTP was spiked with about 10 μ Ci ml⁻¹ [α -³²P] GTP. The extent of GTP hydrolysis at distinct time points during the course of a reaction was analysed by TLC on PEI cellulose (Sigma), and results were quantified using a phosphorimager and ImageQuant software (Molecular Dynamics). We transferred raw data to the GraphPad suite of kinetic analysis programs for determination of the kinetic parameters K_{cat} and K_m .

Preparation of lipid nanotubes

We prepared lipid nanotubes essentially as described^{15,22}. The lipid composition (w/w) was 40% phosphatidylcholine, 40% non-hydroxy fatty acid galactocerebroside, 10% cholesterol and 10% phosphoinositide 4,5-bisphosphate (PtdIns(4,5)P₂). The nanotubes

produced had an average diameter of 28 nm (with little variation) and an average length of 500 nm.

Electron microscopy of dynamin on lipid nanotubes

Samples for electron microscopy were prepared as for GTPase assay reactions in the presence of 100 μ M GTP- γ S or 500 μ M GTP. Samples were applied to glow-discharged 350-mesh grids that were carbon-coated, stained with 0.5% uranyl acetate, and imaged on a Philips 208 80-kV electron microscope.

Thin section electron microscopy of COS- cells

COS-7 cells were transfected with T65A dynamin as for transferrin-uptake assays. Cells were fixed with 2% paraformaldehyde/1.5% glutaraldehyde in 0.1 M sodium cacodylate for 20 min at room temperature. Cells were post-fixed in 1% osmium tetroxide/1.5% potassium ferricyanide and then treated with tannic acid. We then embedded cells as described²³. To test for surface accessibility of tubules, cells were fixed with 2% paraformaldehyde/1.5% glutaraldehyde/0.3% ruthenium red, for 30 min at room temperature. These cells were post-fixed in 1% osmium tetroxide/0.3% ruthenium red, for 1 h at 4 °C, followed by dehydration and embedding. Thin sections were cut on a Reichert-Jung Ultracut E ultramicrotome, stained with lead citrate and viewed using a Phillips CM12 transmission electron microscope.

Received 29 September; accepted 18 January 2001.

- Urrutia, R., Henley, J. R., Cook, T. & McNiven, M. A. The dynamin: redundant or distinct functions for an expanding family of related GTPases? *Proc. Natl Acad. Sci. USA* **94**, 377–384 (1997).
- van der Bliek, A. M. Functional diversity in the dynamin family. *Trends Cell Biol.* **9**, 96–102 (1999).
- Warnock, D. E., Hinshaw, J. E. & Schmid, S. L. Dynamin self-assembly stimulates its GTPase activity. *J. Biol. Chem.* **271**, 22310–22314 (1996).
- Roos, J. & Kelly, R. B. Is dynamin really a 'pinchase'? *Trends Cell Biol.* **7**, 257–259 (1997).
- Sever, S., Muhlberg, A. B. & Schmid, S. L. Impairment of dynamin's GAP domain stimulates receptor-mediated endocytosis. *Nature* **398**, 481–486 (1999).
- Poodry, C. A. & Edgar, L. Reversible alterations in the neuromuscular junctions of *Drosophila melanogaster* bearing a temperature-sensitive mutation, shibire. *J. Cell Biol.* **81**, 520–527 (1979).
- Kosaka, T. & Ikeda, K. Possible temperature-dependent blockage of synaptic vesicle recycling induced by a single gene mutation in *Drosophila*. *J. Neurobiol.* **14**, 207–225 (1983).
- Koenig, J. H. & Ikeda, K. Disappearance and reformation of synaptic vesicle membrane upon transmitter release observed under reversible blockage of membrane retrieval. *J. Neurosci.* **9**, 3844–3860 (1989).
- van der Bliek, A. M. *et al.* Mutations in human dynamin block an intermediate stage in coated vesicle formation. *J. Cell Biol.* **122**, 553–563 (1993).
- Herskovits, J. S., Burgess, C. C., Obar, R. A. & Vallee, R. B. Effects of mutant rat dynamin on endocytosis. *J. Cell Biol.* **122**, 565–578 (1993).
- Damke, H., Baba, T., Warnock, D. E. & Schmid, S. L. Induction of mutant dynamin specifically blocks endocytic coated vesicle formation. *J. Cell Biol.* **127**, 915–934 (1994).
- Takei, K., McPherson, P. S., Schmid, S. L. & DeCamilli, P. Tubular membrane invaginations coated by dynamin rings are induced by GTP- γ S in nerve terminals. *Nature* **374**, 186–190 (1995).
- Switzer, S. M. & Hinshaw, J. E. Dynamin undergoes a GTP-dependent conformational change causing vesiculation. *Cell* **93**, 1021–1029 (1998).
- Takei, K. *et al.* Generation of coated intermediates of clathrin-mediated endocytosis on protein-free liposomes. *Cell* **94**, 131–141 (1998).
- Stowell, M. H. B., Marks, B., Wigge, P. & McMahon, H. T. Nucleotide-dependent conformational changes in dynamin: evidence for a mechanochemical molecular spring. *Nature Cell Biology* **1**, 27–32 (1999).
- Takei, K., Slepnev, V. I., Haucke, V. & De Camilli, P. Functional partnership between amphiphysin and dynamin in clathrin-mediated endocytosis. *Nature Cell Biol.* **1**, 33–39 (1999).
- Muhlberg, A. B., Warnock, D. E. & Schmid, S. L. Domain structure and intramolecular regulation of dynamin GTPase. *EMBO J.* **16**, 6676–6683 (1998).
- Prakash, B., Renault, L., Praefcke, G. J., Herrmann, C. & Wittinghofer, A. Triphosphate structure of guanylate-binding protein 1 and implications for nucleotide binding and GTPase mechanism. *EMBO J.* **19**, 4555–4564 (2000).
- Okamoto, P. M., Tripet, B., Litowski, J., Hodges, R. S. & Vallee, R. B. Multiple distinct coiled-coils are involved in dynamin self-assembly. *J. Biol. Chem.* **274**, 10277–10286 (1999).
- Pai, E. F. *et al.* Refined crystal structure of the triphosphate conformation of H-ras p21 at 1.35 Å resolution: implications for the mechanism of GTP hydrolysis. *EMBO J.* **9**, 2351–2359 (1990).
- Otero, A. D. Transphosphorylation and G protein activation. *Biochem. Pharmacol.* **39**, 1399–1404 (1990).
- Wilson-Kubalek, E. M., Brown, R. E., Celia, H. & Milligan, R. A. Lipid nanotubes as substrates for helical crystallization of macromolecules. *Proc. Natl Acad. Sci. USA* **95**, 8040–8050 (1998).
- Hopkins, C. R. & Trowbridge, I. S. Internalization and processing of transferrin and the transferrin receptor in human carcinoma A431 cells. *J. Cell Biol.* **97**, 508–521 (1983).
- Vallis, Y., Wigge, P., Marks, B., Evans, P. R. & McMahon, H. T. Importance of the pleckstrin homology domain of dynamin in clathrin-mediated endocytosis. *Curr. Biol.* **9**, 257–260 (1999).
- Achiriloaie, M., Barylko, B. & Albanesi, J. P. Essential role of the dynamin pleckstrin homology domain in receptor mediated endocytosis. *Mol. Cell Biol.* **19**, 1410–1415 (1999).
- Smirnova, E., Shurland, D. L., Newman-Smith, E. D., Pishvae, B. & van der Bliek, A. M. A model for dynamin self-assembly based on binding between three different protein domains. *J. Biol. Chem.* **274**, 14942–14947 (1999).
- Grabs, D. *et al.* The SH3 domain of amphiphysin binds the proline-rich domain of dynamin at a single site that defines a new SH3 binding consensus sequence. *J. Biol. Chem.* **272**, 13419–13425 (1997).
- Wigge, P. *et al.* Amphiphysin heterodimers: potential role in clathrin-mediated endocytosis. *Mol. Biol. Cell* **8**, 2003–2015 (1997).

29. Owen, D. J. *et al.* Crystal structure of the Amphiphysin-2 SH3 domain and its role in prevention of dynamin ring formation. *EMBO J.* **17**, 5273–5285 (1998).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

We thank P. Wigge for discussions, M. Ford for discussions and assistance with immunofluorescence work, and M. Higgins for help with the expression of dynamin mutants.

Correspondence and requests for materials should be addressed to H.T.M. (e-mail: hmm@mrc-lmmb.cam.ac.uk).

A mechanism for initiating RNA-dependent RNA polymerization

Sarah J. Butcher*†, Jonathan M. Grimes†‡, Eugeny V. Makeyev*, Dennis H. Bamford* & David I. Stuart‡§

* Institute of Biotechnology and Department of Biosciences, Viikki Biocenter, PO Box 56 (Viikinkaari 5), 00014 University of Helsinki, Finland

‡ Division of Structural Biology, The Henry Wellcome Building for Genomic Medicine, Oxford University, Roosevelt Drive, Oxford OX3 7BN, UK

§ The Oxford Centre for Molecular Sciences, New Chemistry, South Parks Road, Oxford OX1 3QT, UK

† These authors contributed equally to this work

In most RNA viruses, genome replication and transcription are catalysed by a viral RNA-dependent RNA polymerase. Double-stranded RNA viruses perform these operations in a capsid (the polymerase complex), using an enzyme that can read both single- and double-stranded RNA. Structures have been solved for such viral capsids, but they do not resolve the polymerase subunits in any detail^{1,2}. Here we show that the 2 Å resolution X-ray structure of the active polymerase subunit from the double-stranded RNA bacteriophage ϕ 6 (refs 3, 4) is highly similar to that of the polymerase of hepatitis C virus, providing an evolutionary link between double-stranded RNA viruses and flaviviruses. By crystal soaking and co-crystallization, we determined a number of other structures, including complexes with oligonucleotide and/or nucleoside triphosphates (NTPs), that suggest a mechanism by which the incoming double-stranded RNA is opened up to feed the template through to the active site, while the substrates enter by another route. The template strand initially overshoots, locking into a specificity pocket, and then, in the presence of cognate NTPs, reverses to form the initiation complex; this process engages two NTPs, one of which acts with the carboxy-terminal domain of the protein to prime the reaction. Our results provide a working model for the initiation of replication and transcription.

ϕ 6 is amenable to reverse genetics and *in vitro* assembly from purified components⁵, allowing the dissection of both replication (minus-strand synthesis from single-stranded templates) and semi-conservative transcription (plus-strand synthesis from a double-stranded template)^{6,7}. Isolated recombinant ϕ 6 polymerase can use both single-stranded (ss) RNA and double-stranded (ds) RNA as templates with similar elongation rates for the polymerase complex. The rate-limiting step of the transcription reaction is initiation, which occurs at the conserved, terminal 3'-cytidine nucleotide of the negative sense, linear, template RNA^{3,8}.

We determined the structure of recombinant polymerase by multi-wavelength anomalous dispersion of a single crystal of seleno-methionated protein (see Methods). The twofold-averaged

electron density map allowed immediate interpretation of the structure, with all residues placed unambiguously. We also determined the structure of another crystal form with three molecules in the asymmetric unit, providing us with a total of five views of the molecule in different environments.

The structures observed in the two different space groups agree very well (r.m.s. deviation between all C α atoms 0.3 Å), showing that ϕ 6 possesses an unusually rigid polymerase. The 664 residues of the monomeric protein form 25 α -helices and 21 β -strands, which make up the well known canonical hand-like features, with domains representing the fingers, palm and thumb⁹ (Fig. 1a). All the sequence motifs characteristic of RNA polymerases^{10,11} are present (the key aspartic acid residues are 324, 453, 454; see Supplementary Information).

Two elaborations of the basic architecture (Fig. 1a) result in the ϕ 6 polymerase molecule being roughly spherical. First, six strands of polypeptide chain strap together the tips of the fingers and thumb; a four-stranded version of this strap occurs in HCV polymerase^{11–13}. The C-terminal 64 residues form the second elaboration, reminiscent of the C-terminal structure seen in HCV polymerase^{11–13}.

Objective comparison of the polymerases of ϕ 6 and HCV was performed using an automatic procedure (program SHP¹⁴). Breaking the molecules into two rigid bodies allowed superimposition of 418 C α atoms with r.m.s. deviation of 3.5 Å (Fig. 1b) (beyond the established motifs there is essentially no conservation of amino-acid sequence; see Supplementary Information). These two structures resemble each other much more closely than either resembles any other known polymerase structure (Fig. 1b). Automatic superposition places D324, D453 and D454 of the ϕ 6 polymerase within 2 Å of the triad of active-site aspartic acids in HCV polymerase.

The palm domain shows the greatest conservation. There is also notable conservation in the fingers and some similarity in the chains that link the fingers and thumb. The thumb domains must be aligned as a separate rigid body, which then matches 62 residues out of 86 for ϕ 6 (r.m.s. deviation 3.7 Å) with a 35° difference in the relative orientation of these domains (Fig. 1b). The similarity in catalytic side-chain positions, overall molecular architecture and topology, and previously unremarked biological similarities (such as the ability of both polymerases to act without a primer on

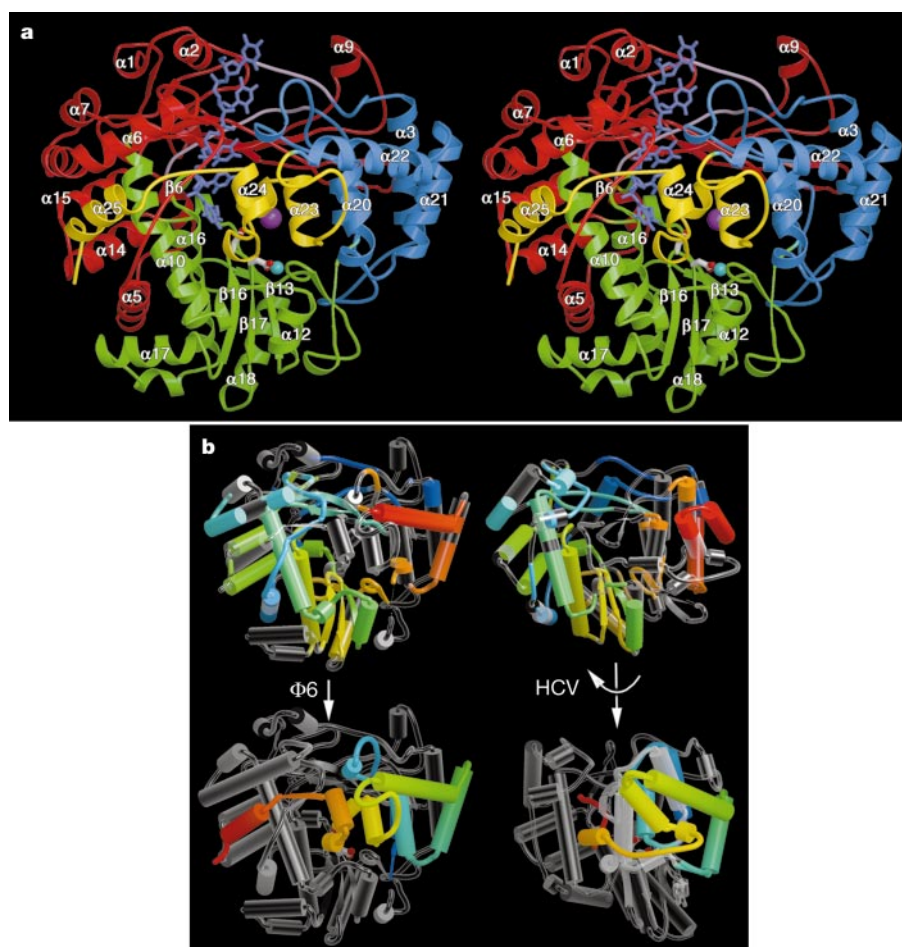


Figure 1 Structure of ϕ 6 polymerase and comparison with HCV polymerase. **a**, Stereo image showing secondary structure elements, coloured according to the generic polymerase domain architecture: red, fingers domain (1–30, 104–276, 333–397); green, palm domain (277–332, 398–517); blue, thumb domain (37–91, 518–600); yellow, C-terminal domain (601–664); mauve, connecting chains (31–36, 92–103). The centre of gravity for the triphosphates of bound NTP is shown as a large purple ball, the bound Mn^{2+} ion is cyan and the bound template is violet. **b**, ϕ 6 and HCV polymerases compared by structural alignment using SHP¹⁴. Top row, ϕ 6 and HCV polymerases. Structurally equivalent stretches of two or more residues are coloured from blue to red according to position in sequence, non-equivalent portions are transparent. For

alignment, amino- and carboxy-terminal portions were superposed separately (the break after the palm domain), matching 418 C α atoms with an r.m.s. deviation of 3.5 Å. Corresponding alignments of poliovirus polymerase and HIV-1 reverse transcriptase versus ϕ 6 polymerase yield 246 C α atoms and 3.8 Å, and 220 C α atoms and 3.6 Å, respectively. Bottom row, thumb and C-terminal domains. HCV polymerase has been rotated (in the direction shown) such that its thumb and C-terminal domains align with the equivalent portions of ϕ 6 polymerase. The polypeptide chain for these domains is coloured as a smooth progression from blue to red at the C terminus. An insert in the HCV structure is shown in light grey, the body of the molecules are transparent.

ssRNA^{3,15}, and stimulation of the polymerases by GTP^{16–18}) indicates that $\phi 6$ and HCV polymerases have diverged from a common ancestor, providing an unexpected evolutionary link between dsRNA viruses and the flaviviruses.

We co-crystallized $\phi 6$ polymerase with an oligonucleotide (TTTCC) that is the DNA equivalent of the consensus sequence of the 3' ends of the minus strands of the $\phi 6$ genome segment (that is, a DNA model of the preferred RNA template). This oligonucleotide binds inside a tunnel lined with predominantly basic amino

acids that leads to the active site, in a position and conformation equivalent to that seen for template bound to HIV-1 reverse transcriptase¹⁹ (Fig. 2a, c), suggesting that the template conformation is functionally important. However, comparison with HIV-1 reverse transcriptase reveals a puzzling feature: mapping the double-stranded DNA product seen for the reverse transcriptase complex onto the $\phi 6$ polymerase structure places it exactly through the body of the C-terminal domain. We can find no way to rebuild the DNA duplex to resolve this conflict. Intriguingly, the same blockade is

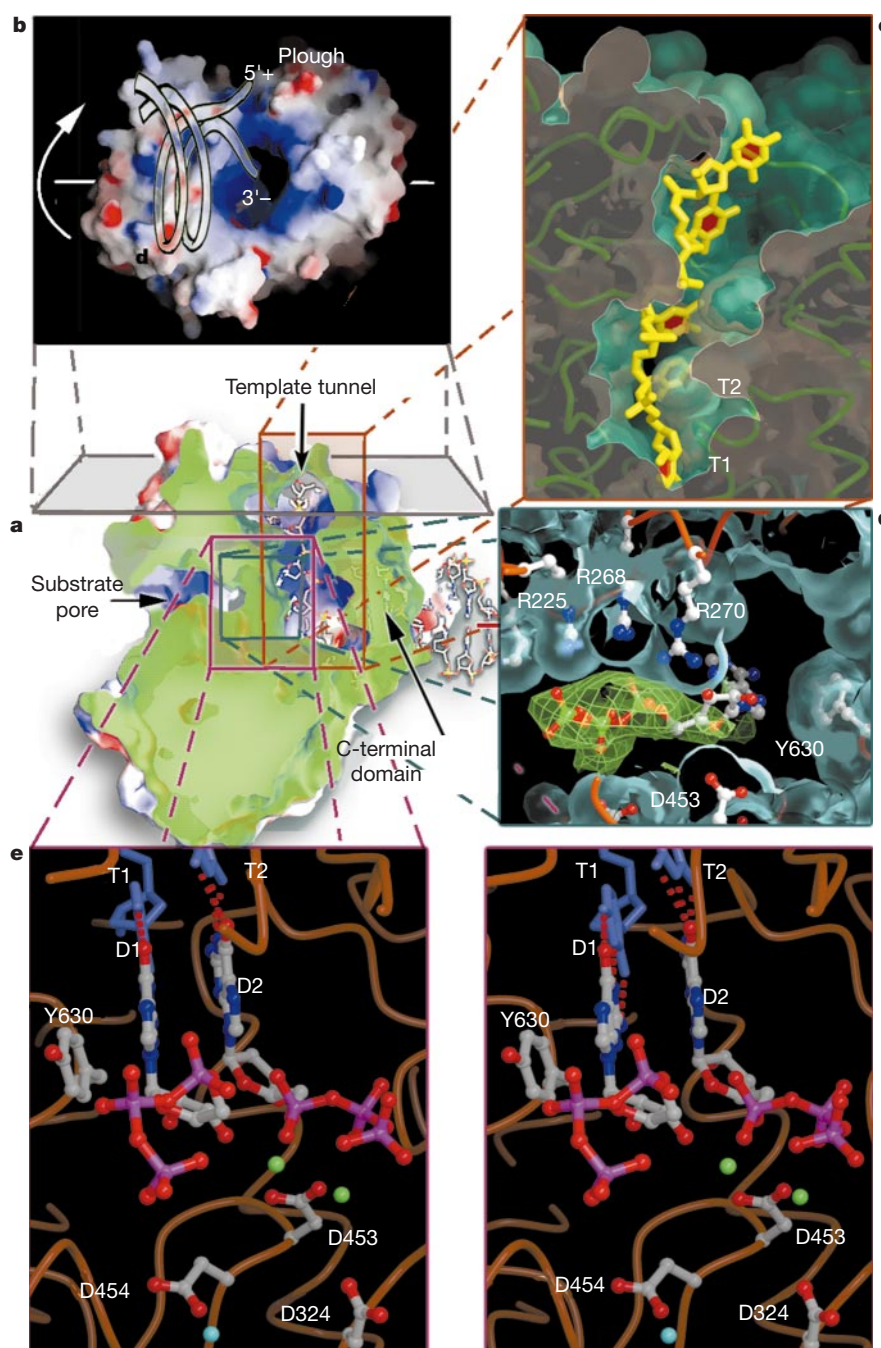


Figure 2 Key aspects of various $\phi 6$ polymerase structures. **a**, $\phi 6$ polymerase sliced open; arrows highlight key features. The areas of the surrounding close-ups in **b–e** are marked by semitransparent boxes, and orientations shown are with respect to **a**.

b, Surface representation²⁸ viewed from above showing the entrance to the template tunnel. Putative positions for the strands before initiation are shown. **c**, A section through the template channel with the bound oligomer drawn in yellow. The surface of the polymerase and the embedded polypeptide chain are coloured green. The two 3'

cytidines are marked as T1 and T2. **d**, Difference electron density map for the NTP bound to site I (based on ninefold averaging of three difference density maps). ATP is drawn with the phosphates coloured green. **e**, Stereo image of the initiation complex. The 3' cytidines (T1 & T2) are drawn in blue. The incoming GTPs, D1 and D2, are shown base paired (bonds in red) to T1 and T2. Y630 ring stacks with the base of D1. D453, D454 and D324 are the catalytic aspartates, the Mn^{2+} ion is shown in cyan, the two catalytic Mg^{2+} ions are in green.

present in HCV polymerase. As both polymerases can self-prime^{3,15}, a plausible hypothesis is that in closing off the active site the C-terminal region may provide a platform for assembly of the initiation complex.

To simplify the discussion we denote the nucleotide at the 3' end of the template T1 and number sequentially from 3' to 5'. Nucleotides of the daughter strand are denoted D1, D2 and so on, such that D1 base pairs with T1. Unexpectedly, the observed template binding places T1 well past the expected catalytic residues, with the base threaded into a specific binding pocket (site S) within the C-terminal domain. In $\phi 6$, transcription is initiated from the 3' end of the minus strand, a strictly conserved cytidine. This specificity is explained by interactions in the S pocket (a hydrogen bond between an OE of E634 and N4 of cytidine would discriminate against the O4 of uracil, and the observed pocket is too small to accommodate either adenine or guanine).

T2 also inserts beyond the catalytic residues, and is base stacked with T3, which seems to be aligned to engage the incoming substrate NTP (we denote site C as the catalytically relevant binding site for incoming NTPs). Sequence analysis indicates a preference for T2 to be a cytidine, although the interactions with the protein that confer this specificity are unclear. Discrimination at T1 and T2 explains how the virus regulates the level of transcription of each genomic strand and promotes the production of plus ssRNA rather than minus ssRNA from the dsRNA template⁸.

$\phi 6$ polymerase can act, without an additional helicase, on blunt-ended dsRNA, showing that it can break open dsRNA and perform successive strand separation reactions⁸. The molecular surface around the entrance to the template tunnel is highly charged, rich in basic residues and seems suited to nonspecific interactions with the phosphate backbone of RNA. We propose that a plough-like protuberance from this molecular surface may separate the strands of RNA, feeding one strand directly into the similarly basic template channel (Fig. 2b) and the other out of the capsid, perhaps through a positively charged surface channel. Selective attachment of the minus strand in the template tunnel ensures that the correct (plus) strand leaves the polymerase complex. Fraying of the double helix to allow this initial recognition would be facilitated by the low melting point of this region of the genome (7 of the first 9 base pairs are A–U).

Crystal soaking experiments using NTPs (Table 1) reveal a binding site in the substrate pore (Fig. 2d, site I) that orders only the triphosphate moieties, by attachment to basic residues K223, R225, R268 and R270. This region of the protein surface is also positively charged in a number of RNA-dependent RNA polymerases (it contains sequence motif F)¹¹. Site I overlaps with, but is displaced by ~ 5 Å from the inferred NTP-binding site in the active site (site C) and might facilitate the interrogation of the template by incoming NTPs.

Both HCV and $\phi 6$ polymerase require divalent cations for activity. The effects of Mg^{2+} and Mn^{2+} are distinct, and, although the effects on HCV are disputed, evidence indicates that in the presence of 10 mM Mg^{2+} both polymerases are activated further by lower concentrations of Mn^{2+} ions, and high Mn^{2+} concentrations inhibit these enzymes²⁰. We determined the structures of the $\phi 6$ polymerase in the presence of either 8 mM Mg^{2+} or 2 mM Mn^{2+} ions (Table 1). By analogy with other polymerases, we would expect divalent cations to act as catalytic ligands of the phosphate groups of the NTP substrate; however, in the absence of NTP, both Mg^{2+} and Mn^{2+} bind at a site ~ 6 Å from the expected catalytic positions. The octahedral coordination in this new site is constructed from a carboxylate oxygen of the catalytic residue D454, an O ϵ atom of E491, the carbonyl oxygen of A495, and three water molecules (see Supplementary Information; and below).

The polymerase–template complex suggested that residue Y630 in the C-terminal domain might provide a platform on which an initiation complex could be constructed. We therefore analysed a polymerase–template co-crystal into which GTP was briefly soaked (GTP being complementary to the cytidines at positions T1 and T2 of the template). The electron density map indicates that we have captured a complete initiation complex, the key features of which are shown in Fig. 2e. The template has ratcheted back by one nucleotide; T1 and T2 are engaged with GTPs, which are poised to become residues D1 and D2 of the daughter strand. D1 forms a Watson–Crick base pair with T1 and is base stacked on Y630 (in other viruses of the same family the equivalent residue is either a tyrosine or a tryptophan; GenBank codes AF261668, AF226851). We denote this site P.

The γ -phosphate of D1 remains intact during the polymerization reaction⁸, and the triphosphate group of D1 is close to, but not coordinating the divalent cation reported above (which in this complex is occupied by Mn^{2+} (present at 2 mM) rather than Mg^{2+} (present at 10 mM)). The side chain of Q629 moves into site S as T1 moves out, allowing the hydrophobic stalk of the side chain to pack under and support the base of T1. Stacked on D1 is the second GTP, D2, base paired to the second nucleotide of the template (T2) and positioned in site C by two Mg^{2+} ions coordinated by catalytic aspartic acids. One ion is in an almost identical position to that reported for the chain-terminated HIV-1 reverse transcriptase structure¹⁹; the other is shifted by several Å from its poorly defined position in the HIV reverse transcriptase structure.

The triphosphates of D2 are stabilized by the same side chains that hold NTPs in nearby site I, in particular R270 and R204, which swing by about 4 Å, allowing the GTP to shuttle between the I and C sites. This may increase the fidelity and stability of the initiation complex. The GTPs are cognate substrates, and the second Mg^{2+} ion is positioned to facilitate attack on the D2 by the 3'-hydroxyl of D1. Furthermore, the bound D1 NTP is displaced from its position seen

Table 1 Data sets and model refinement

	Mg^{2+} /Se-Met	Mn^{2+}	Mn^{2+} /NTP (GTP/AMPPNP/AMPCPP)	Mn^{2+} /oligo	Mn^{2+} / Mg^{2+} /oligo/GTP
Data processing					
Resolution (Å)	50–2.5	30–2.0	20–3.0§	20–3.0	20–3.0
Completeness (%)†	88 (15)	100 (100)	97/100/99	71 (44)	100 (99)
Total data (unique) in thousands	430 (66)	811 (181)	139 (52)/226 (33)/499 (56)	92 (38)	3,213 (55)
Mean I/σ (†)	17.2 (1.4)	21.0 (2.0)	9 (2)/12 (4)/16 (6)	7.4 (1.5)	28.3 (4.9)
R_{merge} (%)†	9.0 (24.0)	5.8 (44.1)	11 (40)/18 (62)/4 (12)	10.0 (43.6)	5.0 (21.2)
Refinement					
R factor (no. of data in thousands)	2-fold NCS	3-fold NCS	3 rigid bodies	3-fold NCS	3-fold NCS
R_{free} (no. of data in thousands)	27.0 (62.5)	22.1 (171.9)	26 (50)/25 (33)/26 (52)	23.7 (52.6)	21.4 (50.8)
No. of non-protein atoms (nucleic acid/ ions/ H_2O)‡	31.0 (3.3)	24.9 (9.1)	26 (3)/29 (2)/25 (3)	27.1 (2.8)	23.9 (2.7)
r.m.s. deviation	–/1/175	–/1/924	–	95/1/–	139/3/–
bond lengths (Å)	0.018	0.016	–	0.013	0.011
bond angles (°)	2.2	1.7	–	1.6	1.5
$\langle B \rangle$ protein/nucleic acid (Å ²)	47/–	45/–	–	51/116	52/113

* All crystals except the Se-Met (which was space group $P3_2$) were space group $P2_1$.

† The numbers in parentheses refer to the highest resolution shell.

‡ In all cases there were 5,265 protein atoms in the model.

§ The AMPPNP data extended only to 3.5 Å resolution.

in the reverse transcriptase complex so that the 3'-hydroxyl is poised to attack the α -phosphate of D1. It seems that we have captured an enzyme-substrate-template initiation complex; however, there is additional density at the distal I site, consistent with this site loosely holding the next incoming nucleoside triphosphate.

Figure 3a shows the likely sequence of events leading to the formation of the initiation complex, derived from our structures (denoted by red frames). Steps I–IV, the template enters the tunnel and is locked in place by interactions with the specificity pocket S. Site I is occupied by NTPs, presumably in rapid exchange. Step V, GTP binds to site P, stabilized by three hydrogen bonds to the cytidine of the template and stacking interactions with Y630. Step VI, the template ratchets backwards, facilitated by electrostatic attraction to R268 and R270, freeing T1 from pocket S. Step VII, a second GTP slips into site P locking the initiation complex into its final form; catalysis releases pyrophosphate, freeing the nucleotide from R268 and R270 so that it can ratchet down, displacing the C-terminal domain of the protein (step VIII). This may be facilitated by attraction of the phosphates of the GTP in site P to the Mn^{2+} ion. The next NTP slips into site C from site I, which sets the ratchet, and chain elongation is underway.

This mechanism explains various features of the system, including the use of triple hydrogen-bonding base pairs at positions 1 and 2 and a weaker double-bond pair at position 3 (preventing locking of the system by binding an ATP at site C before step V). Mn^{2+} at 2 mM and Mg^{2+} at 10 mM segregate into distinct sites, indicating that the pattern of binding may underlie the complex effects reported for these chemically similar ions. The transition from step VII to VIII is supported by further observations: first, the nascent duplex is too bulky to pass outwards through the template tunnel; and second, the C-terminal domain of the polymerase is

more mobile (main chain $\langle B \rangle$, $61 \text{ \AA}^2$) than the rest of the molecule ($\langle B \rangle$, $36 \text{ \AA}^2$), so it seems likely that flexing this domain would occasionally create enough space to allow the base pairs to ratchet down. In particular, residues 601–645 are relatively weakly attached to the body of the polymerase and might hinge up to allow egress of the duplex.

Our proposed mechanism for priming, initiation and elongation, through the base stacking of two nucleotides against a priming aromatic ring and the subsequent movement away of this priming residue to allow egress of product, may be generalized to HCV polymerase.

The strategy of transcription differs between different families of dsRNA viruses: $\phi 6$ (three genome segments) and birnaviruses (two genome segments) act semiconservatively, whereas others including the Reoviridae (with 10–12 genome segments and a larger polymerase molecule) are fully conservative. But there are similarities, such as in the 3' terminal sequences of the minus strands. The conservative strategy requires that the dsRNA emerging from the polymerase must be separated, the newly made plus strand extruded from the viral capsid and the parent minus strand re-annealed with the old plus strand. Two unwindings of dsRNA are required: one on entry to the polymerase; the second to reassort the strands after polymerization. The $\phi 6$ polymerase performs the first of these steps without a helicase; by analogy, therefore, it is possible that the distinct helicase activity identified for members of the family Reoviridae is used instead to drive strand reassortment (Fig. 3b). The observation for one member of the Reoviridae family²¹—that each dsRNA gene segment is attached to a specific transcription complex containing the polymerase subunit—suggests a strategy whereby the 5' end of the plus strand is attached to the polymerase, which is necessarily fully conservative, and would allow extremely efficient re-initiation of transcription. This might explain why only dsRNA viruses with relatively few genome segments use the simpler semiconservative mechanism of transcription. □

Methods

Structure determination

The purification and crystallization of recombinant full-length polymerase, expressed in *Escherichia coli*, are described elsewhere^{3,4}. Two morphologically indistinguishable crystal forms have been analysed, one belonging to space group $P3_2$ and the other to $P2_1$. Crystals were obtained from a cryo-cooled crystal of seleno-methionated protein grown in the presence of 8 mM $MgCl_2$, belonging to space group $P3_2$ ($a = b = 110.1 \text{ \AA}$; $c = 159.4 \text{ \AA}$; $\gamma = 120^\circ$). Cryo-cooling of all polymerase crystals analysed was preceded by briefly rinsing in 25% v/v glycerol. X-ray data were collected as a four wavelength MAD experiment²² on the SBC ID19 beamline at the APS, Argonne, USA. Phasing using 49 Se atoms yielded a high quality electron density map into which a model was built (see Supplementary Information).

Strict twofold non-crystallographic symmetry (NCS) constraints were imposed throughout the refinement procedure²³, to yield a model with R factor 27% and free R factor 31% (see Table 1). The orientations and positions of the three molecules in the asymmetric unit of the $P2_1$ crystal form were found by molecular replacement²⁴. $P2_1$ data were collected from a cryo-cooled crystal grown in the presence of 2 mM $MnCl_2$ ($a = 105.4 \text{ \AA}$, $b = 94.4 \text{ \AA}$, $c = 141.2 \text{ \AA}$, $\beta = 101.3^\circ$) to 2 $\text{\AA}$ resolution on beamline ID-14 EH1, ESRF, Grenoble. Refinement against this data set, using strict threefold NCS constraints, led to a model for the whole molecule (residues 1–664), an Mn^{2+} ion and 994 waters. The current R factor is 22.1%, free R factor 24.9% and 90% of the residues lie in the most favoured regions of the Ramachandran plot²⁵ (Table 1).

Polymerase with substrates and template

Data for three NTP-polymerase complexes, an oligo-polymerase complex and an oligo-GTP-polymerase complex were collected at the ESRF, Grenoble (beamlines ID14-EH4, ID14-EH2 and BM14), at beamline PX9.6, SRS, Daresbury and in-house (Table 1). The NTP-polymerase data sets were collected from crystals that were soaked in either 10 mM AMPNP, AMPCPP or CTP (all in the presence of 2 mM $MnCl_2$). The oligo-polymerase complex data were collected from co-crystals of polymerase and a DNA oligonucleotide (sequence TTTC) in a stoichiometry of 1:1, once again in the presence of 2 mM $MnCl_2$. These co-crystals were also used, after soaking briefly in 10 mM GTP, 10 mM $MgCl_2$ and 2 mM $MnCl_2$ to collect data for the oligo-GTP-polymerase complex.

Weighted difference Fourier maps were calculated²³ using phases derived from rigid-body refinement of the three polymerase molecules (Table 1). These maps were improved by internal threefold averaging. Even after averaging, the three NTP difference maps only showed clear electron density for the triphosphates, and so were cross-averaged. A model for ATP was fitted, by eye, to the electron density and has not been refined (the occupancy

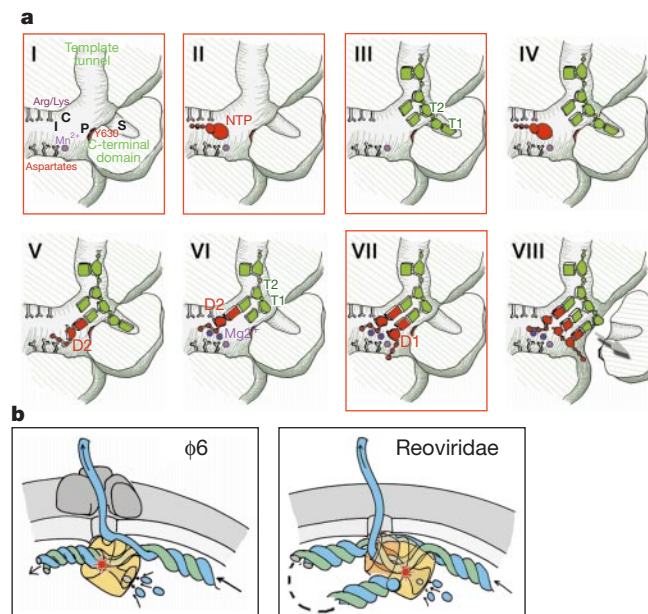


Figure 3 Models for initiation and chain elongation. **a** Cartoon illustrating key points in the reaction mechanism for $\phi 6$ polymerase. Red boxes highlight experimental results. **I**, apo structure with bound Mn^{2+} . Binding sites are identified in black letters. **II**, NTP bound in site I. **III**, template bound. **IV**, template bound and NTP non-productively bound at site I. **V**, initial productive binding at site P. **VI**, template ratchets back. **VII**, second GTP bound at site P. Polymerization can occur. **VIII**, polymerization has occurred, releasing nascent duplex from ordered binding at the active site C. The C-terminal domain moves allowing the duplex to ratchet forward, out of the active site. **b**, $\phi 6$ polymerase and polymerases of the Reoviridae family in the context of the viral capsid. Polymerases are coloured yellow. In the Reoviridae panel the helicase is orange, and the 5' end of the positive strand is attached to the polymerase, holding the genome segment, ready to facilitate re-initiation.

is less than 100%). Models for the additional oligonucleotide, GTP molecules and Mg^{2+} ions, have been fitted into electron density maps and refinement of these oligo-Mn $^{2+}$ -polymerase and oligo-GTP-Mg-Mn-polymerase complexes against their data sets, imposing strict threefold NCS constraints, resulted in models with *R* factors of 23.7 and 21.4%, respectively, and good stereochemistry (Table 1).

Figures

Unless otherwise stated figures were drawn using BOBSCRIPT²⁶ and rendered with RASTER3D²⁷.

Received 2 August; accepted 28 December 2000.

1. Reinisch, K. M., Nibert, M. L. & Harrison, S. C. Structure of the reovirus core at 3.6 Å resolution. *Nature* **404**, 960–967 (2000).
2. Grimes, J. M. *et al.* The atomic structure of the bluetongue virus core. *Nature* **395**, 470–478 (1998).
3. Makeyev, E. V. & Bamford, D. H. Replicase activity of purified recombinant protein P2 of double-stranded RNA bacteriophage φ6. *EMBO J.* **19**, 124–133 (2000).
4. Butcher, S. J., Makeyev, E. V., Grimes, J. M., Stuart, D. I. & Bamford, D. H. Crystallization and preliminary X-ray crystallographic studies on the bacteriophage φ6 RNA-dependent RNA polymerase. *Acta Crystallogr. D* **56**, 1473–1475 (2000).
5. Mindich, L. Reverse genetics of dsRNA bacteriophage φ6. *Adv. Virus Res.* **53**, 341–353 (1999).
6. Gottlieb, P., Strassman, J., Quao, X., Frucht, A. & Mindich, L. In vitro replication, packaging, and transcription of the segmented, double-stranded RNA genome of bacteriophage φ6: studies with procapsids assembled from plasmid-encoded proteins. *J. Bacteriol.* **172**, 5774–5782 (1990).
7. Mindich, L. Precise packaging of the three genomic segments of the double-stranded-RNA bacteriophage φ6. *Microbiol. Mol. Biol. Rev.* **63**, 149–160 (1999).
8. Makeyev, E. V. & Bamford, D. H. The polymerase subunit of a dsRNA virus plays a central role in the regulation of viral RNA metabolism. *EMBO J.* **19**, 124–133 (2000).
9. Ollis, D. L., Kline, C. & Steitz, T. A. Domain of *E. coli* DNA polymerase I showing sequence homology to T7 DNA polymerase. *Nature* **313**, 818–819 (1985).
10. Delarue, M., Poch, O., Tordo, N., Moras, D. & Argos, P. An attempt to unify the structure of polymerases. *Protein Eng.* **3**, 461–467 (1990).
11. Lesburg, C. A. *et al.* Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nature Struct. Biol.* **6**, 937–943 (1999).
12. Ago, H. *et al.* Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Struct. Fold. Des.* **7**, 1417–1426 (1999).
13. Bressanelli, S. *et al.* Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Proc. Natl Acad. Sci. USA* **96**, 13034–13039 (1999).
14. Stuart, D. I., Levine, M., Muirhead, H. & Stammers, D. K. Crystal structure of cat muscle pyruvate kinase at resolution of 2.6 Å. *J. Mol. Biol.* **134**, 109–142 (1979).
15. Oh, J. W., Ito, T. & Lai, M. M. A recombinant hepatitis C virus RNA-dependent RNA polymerase capable of copying the full-length viral RNA. *J. Virol.* **73**, 7694–7702 (1999).
16. Lohmann, V., Overton, H. & Bartenschlager, R. Selective stimulation of hepatitis C virus and pestivirus NS5B RNA polymerase activity by GTP. *J. Biol. Chem.* **274**, 10807–10815 (1999).
17. Frilander, M., Poranen, M. & Bamford, D. H. The large genome segment of dsRNA bacteriophage φ6 is the key regulator in the in vitro minus and plus strand synthesis. *RNA* **1**, 510–518 (1995).
18. van Dijk, A. A., Frilander, M. & Bamford, D. H. Differentiation between minus- and plus-strand synthesis: polymerase activity of dsRNA bacteriophage φ6 in an in vitro packaging and replication system. *Virology* **211**, 320–323 (1995).
19. Huang, H., Chopra, R., Verdine, G. L. & Harrison, S. C. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: implications for drug resistance. *Science* **282**, 1669–1675 (1998).
20. Zhong, W., Uss, A. S., Ferrarri, E., Lau, J. Y. & Hong, Z. De novo initiation of RNA synthesis by hepatitis C virus nonstructural protein 5B polymerase. *J. Virol.* **74**, 2017–2022 (2000).
21. Yazaki, K. & Miura, K. Relation of the structure of cytoplasmic polyhedrosis virus and the synthesis of its messenger RNA. *Virology* **105**, 467–479 (1980).
22. Hendrickson, W. A. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* **254**, 51–58 (1991).
23. Brunger, A. T. *et al.* Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
24. Navaza, J. AMoRe: an automated package for molecular replacement. *Acta Crystallogr. A* **50**, 164–182 (1994).
25. Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
26. Esnouf, R. M. An extensively modified version of MolScript that includes greatly enhanced colouring capabilities. *J. Mol. Graph.* **15**, 132–134 (1997).
27. Merritt, E. A. & Bacon, D. J. in *Macromolecular Crystallography* (eds Carter, J. W. Jr & Sweet, R. M.) 505–524 (Academic, San Diego, 1997).
28. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).

Supplementary information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements

J. Diprose and G. Sutton helped with synchrotron data collection; J. Diprose and S. Ikemizu with calculations; and R. Esnouf and K. Harlos with computing and in-house data collection. We thank the staff at the beamlines of the ESRF, SRS and APS, in particular Sergey Korolev at the APS for help with the MAD experiment. S.J.B. is a Marie Curie Fellow. J.M.G. is funded by the Royal Society and D.I.S. by the Medical Research Council. The work was supported by the Academy of Finland, the Medical Research Council and the European Union.

Correspondence and requests for materials should be addressed to D.I.S. (e-mail: dave@strubi.ox.ac.uk). Coordinates have been deposited in the RCSB Protein database under accession codes: 1HHS, 1HHT, 1HI0, 1HI1, 1HI8.

correction

Improved estimates of global ocean circulation, heat transport and mixing from hydrographic data

Alexandra Ganachaud & Carl Wunsch

Nature **408**, 453–457 (2000).

In this first paragraph of this paper, the uncertainty on the net deep-water production rates in the North Atlantic Ocean was given incorrectly. The correct value should have been $(15 \pm 2) \times 10^6 \text{ m}^3 \text{ s}^{-1}$. □

errata

Changes in Greenland ice sheet elevation attributed primarily to snow accumulation variability

J. R. McConnell, R. J. Arthern, E. Mosley-Thompson, C. H. Davis, R. C. Bales, R. Thomas, J. F. Burkhard & J. D. Kyne

Nature **406**, 877–879 (2000).

As the result of an editing error, the 1993–1998 aircraft-based altimetry surveys of the southern Greenland ice sheet reported by Krabill *et al.* (1999) were erroneously described as satellite-based. □

Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7

Nicole T. Perna, Guy Plunkett III, Valerie Burland, Bob Mau, Jeremy D. Glasner, Debra J. Rose, George F. Mayhew, Peter S. Evans, Jason Gregor, Heather A. Kirkpatrick, György Pósfai, Jeremiah Hackett, Sara Klink, Adam Boutin, Ying Shao, Leslie Miller, Erik J. Grotbeck, N. Wayne Davis, Alex Lim, Eileen T. Dimalanta, Konstantinos D. Potamouis, Jennifer Apodaca, Thomas S. Anantharaman, Jieyi Lin, Galex Yen, David C. Schwartz, Rodney A. Welch & Frederick R. Blattner

Nature **409**, 529–533 (2001).

The Genbank accession number for the annotated sequence given in this paper was typeset incorrectly. The correct accession number is AE005174. □

nature insight

Complex systems



Cover illustration

A computer simulation of the ground displacement due to the 1992 Landers earthquake — an example of one of the many systems that show complex dynamical behaviour. (Image: D. Massonnet/CNES/SPL.)

The science of complexity, as befits its name, lacks a simple definition. It has been used to refer to the study of systems that operate at the 'edge of chaos' (itself a loosely defined concept); to infer structure in the complex properties of systems that are intermediate between perfect order and perfect disorder; or even as a simple restatement of the cliché that the behaviour of some systems as a whole can be more than the sum of their parts.

Notwithstanding these difficulties over formal definition, the study of complex systems has seen tremendous growth. Numerous research programmes, institutes and scientific journals have been established under this banner. And the new concepts emerging from these studies are now influencing disciplines as disparate as astronomy and biology, physics and finance. The richness of the field and the diversity of its application lends itself naturally to the Insight format, although our choice of themes to review is necessarily somewhat eclectic.

We begin by considering systems in which the microscopic properties and processes can be immensely complex and seemingly noisy, yet on larger scales they exhibit certain classes of simple behaviour that seem insensitive to the mechanistic details. On page 242, Sethna *et al.* show that the seemingly random, impulsive events by which many physical systems evolve exhibit universal — and, to some extent, predictable — behaviour. Shinbrot and Muzzio on page 251 offer a different perspective on noise, describing how order and patterns can emerge from intrinsically noisy systems.

We then shift our focus to systems where both the properties of the individual components and the nature of their interactions are reasonably well understood, yet the collective behaviour of the ensemble can still defy simple explanation. On page 259, Debenedetti and Stillingner show how recent theoretical progress on describing the dynamics of systems of many identical interacting particles — in the form of a multidimensional 'energy landscape' — is shedding light on the age-old phenomena of supercooling and glass formation. And for extensive networks of simple interacting systems, Strogatz shows on page 268 how network topology can be as important as the interactions between elements.

But complex systems do not always lend themselves to such easy (if qualitative) categorization. For example, the many complex rhythms encountered in living organisms arise not just from intrinsic stochastic or nonlinear dynamical processes, but also from their interaction with an external fluctuating environment. Yet, according to Glass on page 277, decoding the essential features of these rhythms might ultimately be of benefit to medicine, even in the absence of a simple mathematical interpretation.

As should be clear from these articles, the science of complexity is in its infancy, and some research directions that today seem fruitful might eventually prove to be academic cul-de-sacs. Nevertheless, it seems reasonable to suppose that the general principles emerging from these studies will help us to better understand the complex world around us.

Karl Ziemelis Physical Sciences Editor

Liz Allen Publisher

Crackling noise

James P. Sethna*, Karin A. Dahmen† & Christopher R. Myers‡

*Laboratory of Atomic and Solid State Physics, Clark Hall, Cornell University, Ithaca, New York 14853-2501, USA (sethna@lassp.cornell.edu)

†Department of Physics, 1110 West Green Street, University of Illinois at Urbana-Champaign, Illinois 61801-3080, USA

(dahmen@physics.uiuc.edu)

‡Cornell Theory Center, Frank H. T. Rhodes Hall, Cornell University, Ithaca, New York 14853-3801, USA (myers@tc.cornell.edu)

Crackling noise arises when a system responds to changing external conditions through discrete, impulsive events spanning a broad range of sizes. A wide variety of physical systems exhibiting crackling noise have been studied, from earthquakes on faults to paper crumpling. Because these systems exhibit regular behaviour over a huge range of sizes, their behaviour is likely to be independent of microscopic and macroscopic details, and progress can be made by the use of simple models. The fact that these models and real systems can share the same behaviour on many scales is called universality. We illustrate these ideas by using results for our model of crackling noise in magnets, explaining the use of the renormalization group and scaling collapses, and we highlight some continuing challenges in this still-evolving field.

In the past decade or so, science has broadened its purview to include a new range of phenomena. Using tools developed to understand second-order phase transitions^{1–5} in the 1960s and 70s, stochastic models of turbulence⁶ in the 1970s, and disordered systems^{7–9} in the 1980s, scientists now claim that they should be able to explain how and why things crackle.

Many systems crackle; when pushed slowly, they respond with discrete events of a variety of sizes. The Earth responds¹⁰ with violent and intermittent earthquakes as two tectonic plates rub past one another (Fig. 1). A piece of paper¹¹ (or a candy wrapper at the cinema^{12,13}) emits intermittent, sharp noises as it is slowly crumpled or rumpled. (Try it, but preferably not with this page.) A magnetic material in a changing external field magnetizes in a series of jumps^{14,15}. These individual events span many orders of magnitude in size — indeed, the distribution of sizes forms a power law with no characteristic size scale. In the past few years, scientists have been making rapid progress in developing models and theories for understanding this sort of scale-invariant behaviour in driven, nonlinear, dynamical systems.

Interest in these sorts of phenomena goes back several decades. The work of Gutenberg and Richter¹⁰ in the 1940s and 1950s established the well-known frequency-magnitude relationship for earthquakes that bears their names (Fig. 1). A variety of many-degree-of-freedom dynamical models^{16–28}, with and without disorder, have been introduced in the years since to investigate the nature of slip complexity in earthquakes. More recent impetus for work in this field came from the study of the depinning transition in sliding charge-density wave (CDW) conductors in the 1980s and early 1990s^{29–35}. Interpretation of the CDW depinning transition as a dynamic critical phenomenon sprung from Fisher's early work^{29,30}, and several theoretical and numerical studies followed. This activity culminated in the renormalization-group solution by Narayan and Fisher³² and the numerical studies by Middleton³³ and Myers³⁴, which combined to provide a clear picture of depinning in CDWs and open the doors to the study of other disordered, non-equilibrium systems.

Bak, Tang and Wiesenfeld inspired much of the succeeding work on crackling noise^{36,37}. They introduced the connection between dynamical critical phenomena and crackling noise, and they emphasized how systems may end

up naturally at the critical point through a process of self-organized criticality. (Their original model was that of avalanches in growing sandpiles — sand has long been used as an example of crackling noise^{38,39}, but we now know that real sandpiles do not crackle at the longest scales^{40,41}.)

Researchers have studied many systems that crackle. Simple models have been developed to study bubbles rearranging in foams as they are sheared⁴², biological extinctions⁴³ (where the models are controversial^{44,45} — they ignore catastrophic external events like asteroids), fluids invading porous materials and other problems involving invading fronts^{46–51} (where the model we describe was invented^{46,47}), the dynamics of superconductors^{52–54} and superfluids^{55,56}, sound emitted during martensitic phase transitions⁵⁷, fluctuations in the stock market^{58,59}, solar flares⁶⁰, cascading failures in power grids^{61,62}, failures in systems designed for optimal performance^{63–65}, group decision-making⁶⁶, and fracture in disordered materials^{67–72}. These models are driven systems with many degrees of freedom, which respond to the driving in a series of discrete avalanches spanning a broad range of scales — what in this paper we term crackling noise.

There has been healthy scepticism by some established professionals in these fields to the sometimes-grandiose claims by newcomers claiming support for an overarching paradigm. But often confusion arises because of the unusual kind of predictions the new methods provide. If such models apply at all to a physical system, they should be able to predict most behaviour on long scales of length and time, independent of many microscopic details of the real world. But this predictive capacity comes at a price: the models typically do not make clear predictions of how the real-world microscopic parameters affect the behaviour at long length scales.

Here we provide an overview of the renormalization group^{1–5} used by many researchers to understand crackling noise. Briefly, the renormalization group discusses how the effective evolution laws of a system change as measurements are made on longer and longer length scales. (It works by generating a coarse-graining mapping in system space, the abstract space of possible evolution laws.) The broad range of event sizes are attributed to a self-similarity, where the evolution laws look the same at different length scales. This self-similarity leads to a method for scaling experimental data. In the simplest case this yields power laws and fractal

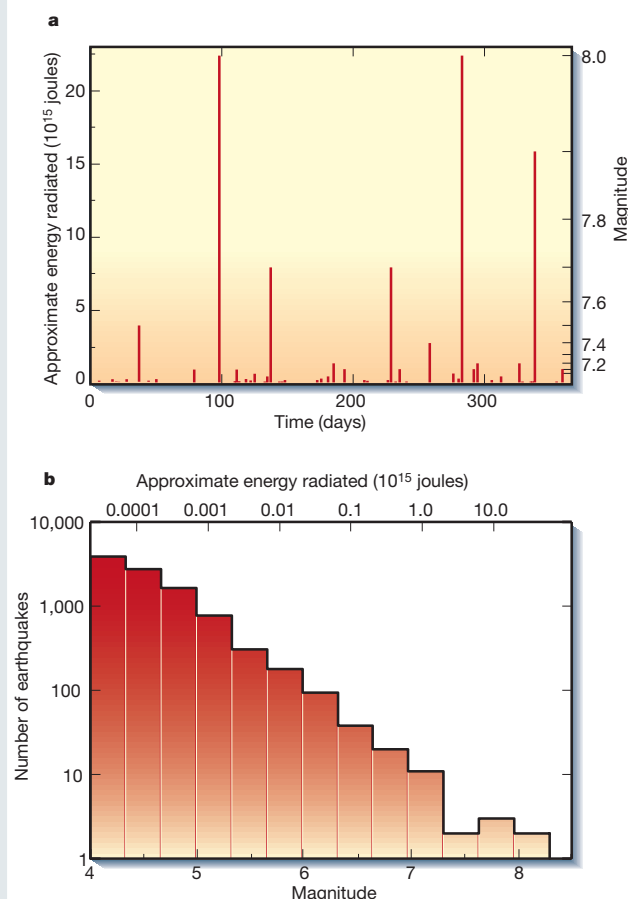


Figure 1 The Earth cracks. **a**, Time history of radiated energy from earthquakes throughout all of 1995^{108–110}. The Earth responds to the slow strains imposed by continental drift through a series of earthquakes (impulsive events well separated in space and time). This time series, when sped up, sounds remarkably like the crackling noise of paper, magnets and Rice Krispies (listen to it in ref. 110). **b**, Histogram of number of earthquakes in 1995 as function of their magnitude (or, alternatively, their energy release). Earthquakes come in a wide range of sizes, from unnoticeable trembles to catastrophic events. The smaller earthquakes are much more common: the number of events of a given size forms a power law¹⁰ called the Gutenberg–Richter law. (Earthquake magnitude scales with the logarithm of the strength of the earthquake. On a log–log plot of number versus radiated energy, the power law is a straight line, as we observe in the plotted histogram.) One would hope that such a simple law should have an elegant explanation.

structures, but more generally it leads to universal scaling functions — where we argue the real predictive power lies. We will only touch upon the complex analytical methods used in this field, but we believe we can explain faithfully and fully both what our tools are useful for, and how to apply them in practice. The renormalization group is perhaps the most impressive use of abstraction in science.

Why should crackling noise be comprehensible?

Not all systems crackle. Some respond to external forces with many similar-sized, small events (for example, popcorn popping as it is heated). Others give way in one single event (for example, chalk snapping as it is stressed). In broad terms, crackling noise is in between these limits: when the connections between parts of the system are stronger than in popcorn but weaker than in the grains making up chalk, the yielding events can span many size scales. Crackling forms the transition between snapping and popping.

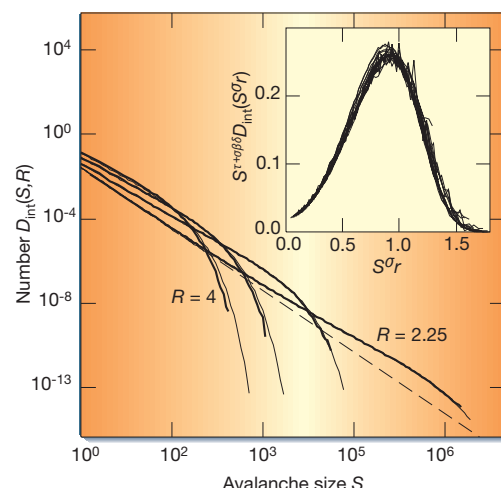


Figure 2 Magnets crackle^{73–76}. Magnets respond to a slowly varying external field by changing their magnetization in a series of bursts, or avalanches. These bursts, called Barkhausen noise, are very similar (albeit on different time and size scales) to those shown in Fig. 1 for earthquakes. The avalanches in our model have a power-law distribution only at a special value of the disorder, $R_c = 2.16$. Shown is a histogram giving the number of avalanches $D_{\text{int}}(S, R)$ of a given size S at various disorders R ranging from 4 to 2.25; the thin lines are theoretical predictions from our model. (D_{int} gives all the avalanches during our simulation, integrated over the external field $-\infty < H(t) < +\infty$). The straight dashed line shows the power-law distribution at the critical point. Notice that fitting power laws to the data would work only very near to R_c : even with a range of a million in avalanche sizes, the slope has not converged to the asymptotic value. On the other hand, scaling-function predictions (theoretical curves) work well far from the critical point. The inset shows a scaling collapse of the avalanche size distribution (scaled probability versus scaled size), which is used to provide the theoretical curves as described in the text.

Figure 1b presents a simple relationship between earthquake number and magnitude. We expect that there ought to be a simple, underlying reason why earthquakes occur on all different sizes. The properties of very small earthquakes probably depend in detail on the kind of dirt (fault gouge) in the crack. The very largest earthquakes will depend on the geography of the continental plates. But the smooth power-law behaviour indicates that something simpler is happening in between, independent of either the microscopic or the macroscopic details.

There is an analogy here with the behaviour of a fluid. A fluid is very complicated on the microscopic scale, where molecules are bumping into one another: the trajectories of the molecules are chaotic, and depend both on exactly what direction they are moving and what they are made of. However, a simple law describes most fluids on long time and size scales. This law, the Navier–Stokes equation, depends on the constituent molecules only through a few parameters (the density and viscosity). Physics works because simple laws emerge on large scales. In fluids, these microscopic fluctuations and complexities disappear on large scales: for crackling noise, they become scale-invariant and self-similar.

How do we derive the laws for crackling noise? There are two approaches. First, we can calculate analytically the behaviour on long time and size scales by formally coarse-graining over the microscopic fluctuations. This leads us to renormalization-group methods^{1–5}, which we discuss in the next section. The analytic approach can be challenging, but it can give useful results and (more important) is the only explanation for why events on all scales should occur. Second, we can make use of universality. If the microscopic details do not matter for the behaviour at long length scales, why not make up a simple model with the same behaviour (in the same universality class) and solve it?

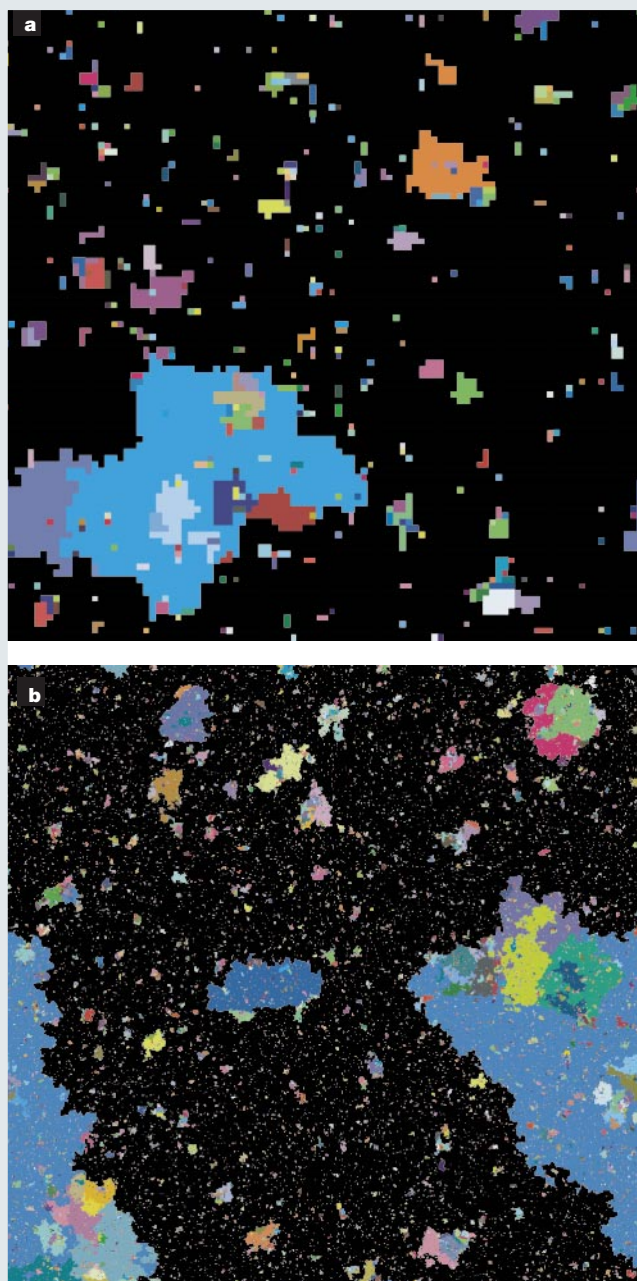


Figure 3 Self-similarity. Cross-sections of the avalanches during the magnetization of our model^{15,73–76}. Here each avalanche is drawn in a separate colour. **a**, A 100^3 simulation; **b**, a $1,000^3$ simulation (a billion domains⁷⁴); both are run at the critical point $R_c = 2.16$ J where avalanches just barely continue. The black background represents a large avalanche that spans the system: the critical point occurs when avalanches would first span an infinite system.

The model we focus on here is a caricature of a magnetic material^{15,46,47,73–77}. A piece of iron will ‘crackle’ as it enters a strong magnetic field, giving what is called Barkhausen noise. We model the iron as a cubic grid of magnetic domains S_i , whose north pole is either pointing upwards ($S_i = +1$) or downwards ($S_i = -1$). The external field pushes on our domain with a force $H(t)$, which will increase with time. Iron can be magnetized because neighbouring domains prefer to point in the same direction: if the six neighbours of our cubic domain are S_j , then in our model we let their force on our domain be $\sum_j JS_j$ (where we set the coupling $J=1$). Finally, we model dirt, randomness in the domain shapes, and other kinds of disorder by introducing a random field h_i , different for each domain and chosen

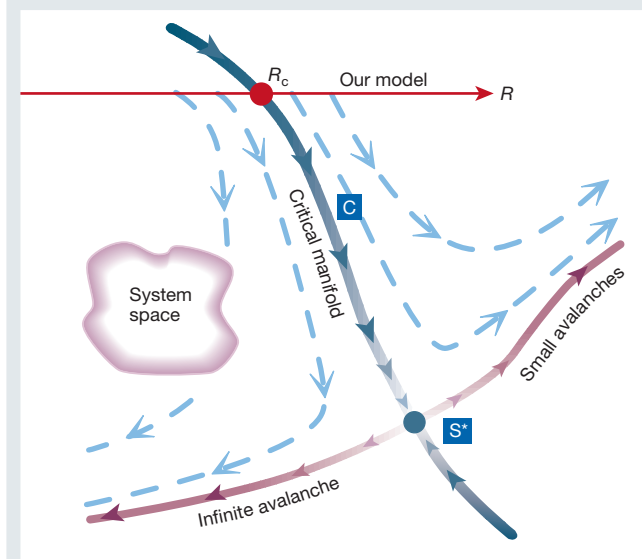


Figure 4 Renormalization-group flows. The renormalization group is a theory of how coarse-graining to longer length scales introduces a mapping from the space of physical systems to itself. Consider the space of all possible models of magnetic hysteresis. Each model can be coarse-grained, removing some fraction of the microscopic degrees of freedom and introducing more complicated rules so that the remaining ones still flip at the same external fields. This defines a mapping from our space into itself. A fixed point S^* in this space will be self-similar: because it maps to itself upon coarse-graining, it must have the same behaviour on different length scales. Points that flow into S^* under coarse-graining share this self-similar behaviour on sufficiently long length scales: they all share the same universality class.

at random from a normal distribution with standard deviation R , which we call the disorder. The net force on our domain is thus

$$\text{Force on domain } i = H(t) + \sum_j JS_j + h_i \quad (1)$$

The domains in our model all start with their north pole pointing down (-1), and flip up as soon as the net force on them becomes positive. This can occur either because $H(t)$ increases sufficiently (spawning a new avalanche), or because one of their neighbours flipped up, kicking them over (propagating an existing avalanche). (Thermal fluctuations are ignored: a good approximation in many experiments because the domains are large.) If the disorder R is large, so the h_i are typically big compared to J , then most domains flip independently: all the avalanches are small, and we get popping noise. If the disorder is small compared to J , then typically most of the domains will be triggered by one of their neighbours: one large avalanche will snap up most of our system. In between, we get crackling noise. When the disorder R is just large enough so that each domain flip on average triggers one of its neighbours (at the critical disorder R_c), then we find avalanches on all scales (Figs 2, 3).

What do these avalanches represent? In nonlinear systems with many degrees of freedom, there are often large numbers of metastable states. Local regions in the system can have multiple stable configurations, and many combinations of these local configurations are possible. (A state is metastable when it cannot lower its energy by small rearrangements. It is distinguished from the globally stable state, which is the absolute lowest energy possible for the system.) Avalanches are the rearrangements that occur as our system shifts from one metastable state to another. Our specific interest is in systems with a broad distribution of avalanche sizes, where shifting between metastable states can rearrange anything between a few domains and millions of domains.

There are many choices we made in our model that do not matter at long time and size scales. Because of universality, we can argue^{78,79}

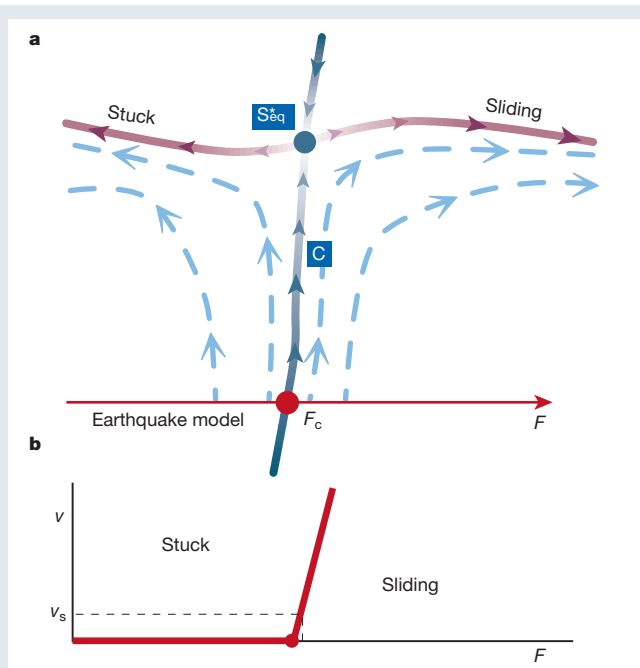


Figure 5 Flows in the space of earthquake models. A model for earthquakes will have a force F applied across the front. In models ignoring inertia and velocity-dependent friction²⁸, there is a critical force F_c that just allows the fault to slip forward. **a**, Coarse-graining defines a flow on the space of earthquake models. The fixed point S^*_{eq} will have a different local flow field from other renormalization-group fixed points, yielding its own universality class of critical exponents and scaling functions. The critical manifold C , consisting of models that flow into S^*_{eq} , separates the stuck faults from those that slide forward with an average velocity $v(F)$. **b**, The velocity varies with the external force as a power law $v(F) \sim F^\beta$. The motion of the continental plates, however, does not fix the force F across the fault: rather, it sets the average relative velocity to a small value v_s (centimetres per year). This automatically sets the force across the fault very close to its critical force F_c . This is one example of self-organized criticality^{36,37}.

that the behaviour would be the same if we chose a different grid of domains, or if we changed the distribution of random fields, or if we introduced more realistic random anisotropies and random coupling constants. Were this not the case, we could hardly expect our simple model to explain real experiments.

The renormalization group and scaling

To study crackling noise, we use renormalization-group^{1–5,78,80} tools developed in the study of second-order phase transitions. The word renormalization has roots in the study of quantum electrodynamics, where the effective charge changes in size (norm) as a function of length scale. The word group refers to the family of coarse-graining operations that are basic to the method: the group product is composition (coarsening repeatedly). The name is unfortunate, however, as the basic coarse-graining operation does not have an inverse, so that the renormalization group does not have the mathematical structure of a group.

The renormalization group studies the way the space of all physical systems maps into itself under coarse-graining (Fig. 4). The coarse-graining operation shrinks the system and removes degrees of freedom on short length scales. Under coarse-graining, we often find a fixed point S^* : many different models flow into the fixed point and hence share long-wavelength properties. Figure 3 provides a schematic view of coarse-graining: the $1,000^3$ cross-section looks (statistically) like the 100^3 section if you blur your eyes by a factor of ten. Much of the mathematical complexity of this field involves finding analytical tools for computing the flow diagram in Fig. 4. Using methods developed to study thermodynamical phase

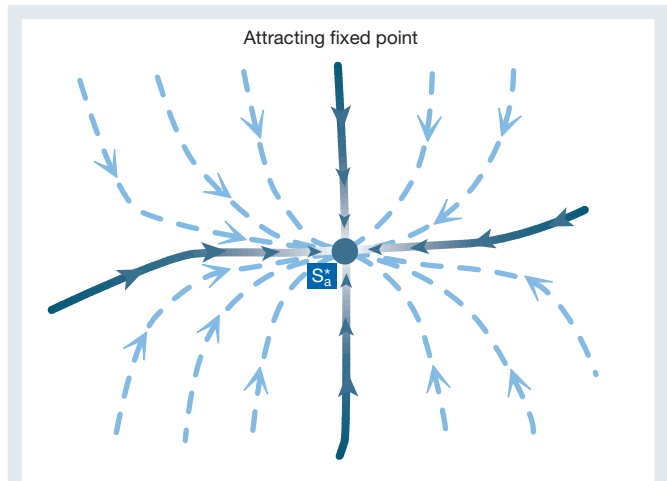


Figure 6 Attracting fixed point. Often there will be fixed points that attract in all directions. These fixed points describe phases rather than phase transitions. Most phases are rather simple, with fluctuations that die away on long length scales⁸¹. When fluctuations remain important, they will exhibit self-similarity and power laws called generic scale invariance^{83,84}.

transitions² and the depinning of charge-density waves³², we can calculate for our model the flows for systems in dimensions close to six (the so-called ϵ -expansion^{78–80}, where $\epsilon = 6 - d$, with d being the dimension of the system). Interpolating between dimensions may seem a surprising thing to do. In our system it gives good predictions even in three dimensions (that is, $\epsilon = 3$), but it is difficult and is not discussed here. Nor will we discuss real-space renormalization-group methods¹ or series-expansion methods. We focus on the relatively simple task of using the renormalization group to justify and explain the universality, self-similarity and scaling observed in nature.

Consider the 'system space' for disordered magnets. There is a separate dimension in system space for each possible parameter in a theoretical model (disorder, coupling, next-neighbour coupling, dipolar fields, and so on) or in an experiment (for example, temperature, annealing time and chemical composition). Coarse-graining, however one implements it, gives a mapping from system space into itself: shrinking the system and ignoring the shortest length scales yields a new physical system with identical long-distance physics, but with different (renormalized) values of the parameters. We have abstracted the problem of understanding crackling noise in magnets into understanding a dynamical system acting on a space of dynamical systems.

Figure 4 represents a two-dimensional cross-section of this infinite-dimensional system space. We have chosen the cross-section to include our model (equation (1)): as we vary the disorder R , our model sweeps out a straight line (red) in system space. The cross-section also includes a fixed point S^* , which maps into itself under coarse-graining. The system S^* looks the same on all scales of length and time, because it coarse-grains into itself. We can picture the cross-section of Fig. 4 either as a plane in system space (in which case the arrows and flows depict projections, as in general the real flows will point somewhat out of the plane), or as the curved manifold swept out by our one-parameter model as we coarse-grain (in which case the flows above our model and below the maroon curved line should be ignored).

The flow near S^* has one unstable direction, leading outwards along the maroon curve (the unstable manifold). In system space, there is a surface of points C which flow into S^* under coarse-graining. Because S^* has only one unstable direction, C divides system space into two phases. To the left of C , the systems will have one large, system-spanning avalanche (a snapping noise). To the

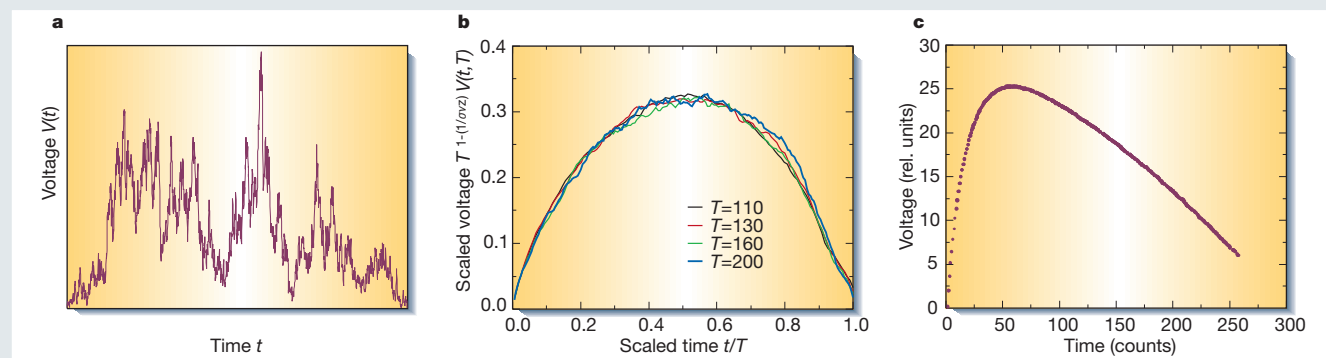


Figure 7 Scaling of avalanche shapes. **a**, Voltage pulse (number of domains flipped per unit time) during a single large avalanche (arbitrary units). Notice how the avalanche almost stops several times: if the forcing were slightly smaller, this large avalanche would have broken up into two or three smaller ones. The fact that the forcing is just large enough to keep the avalanche growing is the cause of the self-similarity: on average a partial avalanche of size S will trigger one other of size S . **b**, Average avalanche shapes⁹⁴ for avalanches of different duration for our model (A. Mehta and K.A.D., unpublished results). In addition to predicting power laws,

our theories should describe all behaviour on long scales of length and time (at least in a statistical sense). In particular, by fixing parameters one can predict what are called scaling functions. If we average the voltage as a function of time over all avalanches of a fixed duration, we obtain an average shape. In our simulation, this shape is the same for different durations. **c**, Experimental data of Spasojević *et al.*⁹⁵ showing all large avalanches averaged after scaling to fixed duration and area. The experimental average shape is very asymmetric and is not described correctly by our model.

right of **C**, all avalanches are finite and under coarse-graining they all become small (popping noise). As it crosses **C** at the value R_c , our model goes through a phase transition.

Our model at R_c is not self-similar on the shortest length scales (where the square lattice of domains still is important), but because it flows into S^* as we coarse-grain we deduce that it is self-similar on long length scales. Some phase transitions, such as ice melting into water, are abrupt and do not exhibit self-similarity. Continuous phase transitions like ours almost always have self-similar fluctuations on long length scales. In addition, note that our model at R_c will have the same self-similar structure as S^* does. Indeed, any experimental or theoretical model lying on the critical surface **C** will share the same long-wavelength critical behaviour. This is the fundamental explanation for universality.

The flows in system space can vary from one class of problems to another: the system space for some earthquake models (Fig. 5a) will have a different flow, and its fixed point will have different scaling behaviour (yielding a different universality class). In some cases, a fixed point will attract all the systems in its vicinity (no unstable directions; Fig. 6). Usually at such attracting fixed points the fluctuations become unimportant at long length scales: the Navier–Stokes equation for fluids described earlier can be viewed as a stable fixed point^{81,82}. The coarse-graining process, averaging over many degrees of freedom, naturally smoothens out fluctuations, if they are not amplified near a critical point by the unstable direction. Fluctuations can remain important when a system has random noise in a conserved property, so that fluctuations can die away only by diffusion: in these cases, the whole phase will have self-similar fluctuations, leading to generic scale invariance^{83,84}.

Sometimes, even when the system space has an unstable direction as in Fig. 4, the observed behaviour always has avalanches of all scales. This can occur simply because the physical system averages over a range of model parameters (that is, averaging over a range of R including R_c in Fig. 4). For example, this can occur by the sweeping of a parameter⁸⁵ slowly in time, or varying it gradually in space — either deliberately or through large-scale inhomogeneities.

Self-organized criticality^{36,37} can also occur, where the system is controlled so that it sits naturally on the critical surface. Self-organization to the critical point can occur through many mechanisms. In some models of earthquake faults (Fig. 5b), the external force naturally stays near the rupture point because the plates move at a fixed, but very small²⁰, velocity with respect to one another (Fig. 5b). (This probably does not occur during large earthquakes, where

inertial effects lead to temporary strain relief^{28,86}.) Sandpile models self-organize (if sand is added to the system at an infinitesimal rate) when open boundary conditions⁸⁷ are used (which allows sand to leave until the sandpile slope falls to the critical value). Long-range interactions^{88–90} between domains can act as a negative feedback in some models, yielding a net external field that remains at the critical point. For each of these cases, once the critical point is understood, adding the mechanism for self-organization is relatively easy.

The case shown in Fig. 4 of ‘plain old criticality’ is what is seen in some^{15,73–76} but not all^{88–92} models of magnetic materials, in foams⁴², and in some models of earthquakes²⁸.

Beyond power laws

The renormalization group is the theoretical basis for understanding why universality and self-similarity occur. Once we accept that different systems should sometimes share long-distance properties, though, we can quite easily derive some powerful predictions.

To take a tangible example, consider the relation between the duration of an avalanche and its size. In paper crumpling, this is not interesting: all the avalanches seem to be without internal temporal structure¹¹. But in magnets, large events take longer to finish, and have an interesting internal statistical self-similarity (Fig. 7a). If we look at all avalanches of a certain duration T in an experiment, they will have a distribution of sizes S around some average $\langle S \rangle_{\text{experiment}}(T)$. If we look at a theoretical model, it will have a corresponding average size $\langle S \rangle_{\text{theory}}(T)$. If our model describes the experiment, these functions must be essentially the same at large S and large T . We must allow for the fact that the experimental units of time and size will be different from the ones in the model: the best we can hope for is that $\langle S \rangle_{\text{experiment}}(T) = A \langle S \rangle_{\text{theory}}(T/B)$, for some rescaling factors A and B .

Now, instead of comparing our model to experiment, we can compare it to itself on a slightly larger timescale⁹³. If the timescale is expanded by a small factor $B = 1/(1 - \delta)$, then the rescaling of the size will also be small, say $1 + a\delta$. Now

$$\langle S \rangle(T) = (1 + a\delta) \langle S \rangle((1 - \delta)T) \quad (2)$$

Making δ very small yields the simple relation $a \langle S \rangle = T d \langle S \rangle / dT$, which can be solved to give the power-law relation $\langle S \rangle(T) = S_0 T^a$. The exponent a is called a critical exponent, and is a universal prediction of a given theory. (This means that if the theory correctly describes an experiment, the critical exponents will agree.) In our

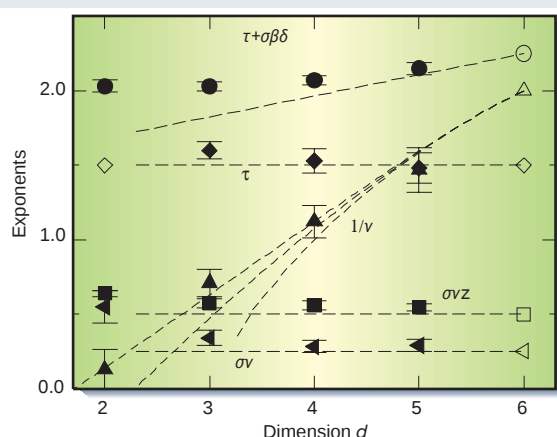


Figure 8 Critical exponents in various dimensions. We test our ε -expansion predictions^{78,80} by measuring⁷⁶ the various critical exponents numerically in up to five spatial dimensions. The various exponents are described in the text. All exponents are calculated only to linear order in ε , except for the correlation length exponent ν , where we use results from other models^{78,80}. The agreement even in three dimensions is remarkably good, considering that we are expanding in ε where $\varepsilon = 3$. Note that perturbing in dimension for our system is not only complicated, but also controversial¹¹¹ (see also section VI.C of ref. 78 and section V of ref. 76).

work, we write the exponent a relating time to size in terms of three other critical exponents, $a = 1/\sigma\nu$.

There are several basic critical exponents, which arise in various combinations depending on the physical property being studied. The details of the naming and relationships between these exponents are not a focus of this article. Briefly, the cutoff in the avalanche size distribution in Fig. 2 gets larger as one approaches the critical disorder as $(R - R_c)^{-1/\sigma}$ (Fig. 2). The typical length L of the largest avalanche is proportional to $(R - R_c)^{-\nu}$. At R_c , the probability of having an avalanche of size S is $S^{-(\tau + \sigma\beta\delta)}$ (Fig. 2); and just at the critical field it is $S^{-\tau}$. (Note that the small change in scale δ should not be confused with the critical exponent δ .) The fractal dimension of the avalanches is $1/\sigma\nu$, meaning the spatial extent L of an avalanche is proportional to the size $S^{\sigma\nu}$. The duration T of an avalanche of spatial extent L is L^z .

To specialists in critical phenomena, these exponents are central; whole conversations will seem to rotate around various combinations of Greek letters. Critical exponents are one of the relatively easy parameters to calculate from the various analytic approaches, and so have attracted the most attention. They are derived from the eigenvalues of the linearized flows about the fixed point S^* in Fig. 4. Figure 8 shows our numerical estimates⁷⁶ for several critical exponents in our model in various spatial dimensions, together with our 6- ε expansions^{78,80} for them. Of course the key challenge is not to get analytical work to agree with numerics: it is to get theory to agree with experiment. Figure 9 shows that our model does well in describing a wide variety of experiments, but that two alternative models (with different flows around their fixed points) also fit.

Critical exponents are not everything; many other scaling predictions, explaining wide varieties of behaviour, are relatively straightforward to extract from numerical simulations. Universality extends even to those properties of long length scales for which written formulas do not yet exist. Perhaps the most important of these other predictions are the universal scaling functions. For example, consider the time history of the avalanches, $V(t)$, denoting the number of domains flipping per unit time. (We call it V because it is usually measured as a voltage in a pickup coil.) Each avalanche has large fluctuations, but we can average over many avalanches to get a typical shape. Figure 7b shows the average over all avalanches of fixed duration T , which we shall call $\langle V \rangle(T, t)$. Universality again suggests that this average should be the same for experiment and a successful

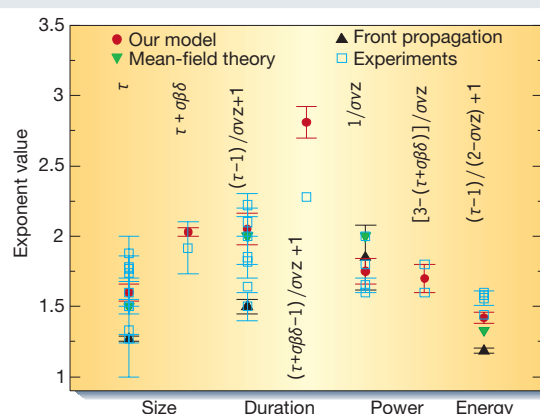


Figure 9 Comparing experiments with theory: critical exponents. Different experiments on crackling noise in magnets measure different combinations of the universal critical exponents. Here we compare experimental measurements^{88,95,112–122} (see also Table I of ref. 73) to the theoretical predictions for three models: our model^{15,73–76}, the front-propagation model^{46–51,121} and mean-field theory. (In mean-field theory our coupling J in equation (1) couples all pairs of spins: such long-range interactions occur because of the boundaries in models with magnetic dipolar forces⁹⁰. Mean-field theory is equivalent to a successful model with a single degree of freedom^{123,124}.) Shown are power laws that give the probability of obtaining an avalanche of a given size, duration or energy at the critical point; also shown is the critical exponent giving the power as a function of frequency⁹⁴ (due to the internal structure of the avalanches; Fig. 7a). In each pair of columns, the first column includes only avalanches at external fields H in equation (1) where the largest avalanches occur, and the second column (when it exists) includes all avalanches. The various combinations of the basic critical exponents can be derived from exponent equality calculations similar to the one discussed in the text^{73,80,94}. Many of the experiments were done years before the theories were developed, and many did not report error bars. All three theories do well (especially considering the possible systematic errors in fitting power laws to the experimental measurements; see Fig. 2). Recent work indicates a clumping of experimental values around the mean-field and front-propagation predictions¹²¹.

theory, apart from an overall shift in time and voltage scales: $\langle V \rangle_{\text{experiment}}(T, t) = A \langle V \rangle_{\text{theory}}(T/B, t/B)$. Comparing our model to itself with a shifted timescale becomes straightforward if we change variables: let $v(T, t/T) = \langle V \rangle(T, t)$, so $v(T, t/T) = Av(T/B, t/T)$. Here t/T is a particularly simple example of a scaling variable. Now, if we rescale time by a small factor $B = 1/(1 - \delta)$, we have $v(T, t/T) = (1 + b\delta)v(t/T, (1 - \delta)t/T)$. Again, making δ small we find $bv = T\partial v/\partial T$, with solution $v = v_0 T^b$. However, the integration constant v_0 will now depend on t/T , $v_0 = V(t/T)$, so we arrive at the scaling form

$$\langle V \rangle(t, T) = T^b \mathbf{V}(t/T) \quad (3)$$

where the entire scaling function $\mathbf{V}$ is a universal prediction of the theory.

Figure 7b,c shows the universal scaling functions $\mathbf{V}$ for our model⁹⁴ and an experiment⁹⁵. For our model, we have drawn what are called scaling collapses, a simple but powerful way both to check that we are in the scaling regime, and to measure the universal scaling function. Using the form of the scaling equation (3), we simply plot $T^{-b} \langle V \rangle(t, T)$ versus t/T , for a series of long times T . All the plots fall onto the same curve, which means that our avalanches are large enough to be self-similar. (If in the scaling collapse the corresponding plots do not all look alike, then any power laws measured are probably accidental.) The scaling collapse also provides us with a numerical evaluation of the scaling function $\mathbf{V}$. Note that we use $1/\sigma\nu - 1$ for the critical exponent b . This is an example of an

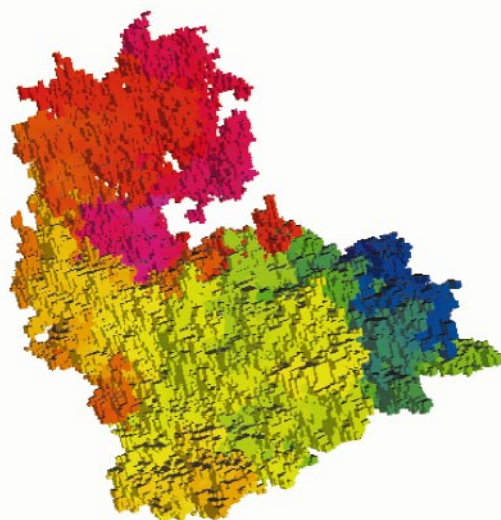


Figure 10 Fractal spatial structure of an avalanche⁷⁴. Fractal structures, as well as power laws, are characteristic of systems at their critical point. This moderate-sized avalanche involved the flipping of 282,785 domains in our simulation. The colours represent time: the first domains to flip are coloured blue, and the last pink. So far, there have not been many experiments showing the spatial structure of avalanches¹²⁵. When experiments become available, there are a wealth of predictions of the scaling theories that we could test. Other systems^{28,55,126} display a qualitatively different kind of avalanche spatial structure, where the avalanche is made up of many small disconnected pieces, which trigger one another through the waves emitted as they flip.

exponent equality and is easily derived from the fact that $\langle S \rangle(T) = \int \langle V \rangle(t, T) dt = \int T^b V(t/T) dt \sim T^{b+1}$, and the scaling relation $\langle S \rangle(T) \sim T^{1/\nu}$.

Notice that our model and the experiment have different shapes for V . The other two models from Fig. 9 also give much more symmetrical forms for V than the experiment does⁹⁴. How do we react to this? Our models are falsified if any of the predictions are shown to be wrong asymptotically on long scales of length and time. If duplication of this measurement by other groups continues to show this asymmetry, then our theory is obviously incomplete. Even if later experiments in other systems agree with our predictions, it would seem that this particular system is described by an undiscovered universality class. Incorporating insights from careful experiments to refine the theoretical models has always been crucial in the broad field of critical phenomena. The message we emphasize here is that scaling functions can provide a sharper tool for discriminating between different universality classes than critical exponents.

In broad terms, most common properties that involve large scales of length and time have scaling forms. Using self-similarity, we can write functions of N variables in terms of scaling functions of $N-1$ variables: $F(x, y, z) = z^{-\alpha} F(x/z^\beta, y/z^\gamma)$. In the inset to Fig. 2, we show the scaling collapse for the avalanche size distribution: $D(S, R) = S^{-(\tau + \alpha\beta\delta)} D((R - R_c)/S^{-\sigma})$. (This example illustrates that scaling works not only at R_c but also near R_c ; the maroon unstable manifold in Fig. 4 governs the behaviour for systems near the critical manifold C .)

Many other kinds of properties beyond critical exponents and scaling functions can be predicted from these theories. Figure 10 shows the spatial structure of a large avalanche in our model: notice that not only is it fractal (rugged on all scales), but also that it is longer than it is wide⁹⁶, and that it is topologically interesting⁹⁷. (It has tunnels, and sometimes during the avalanche it forms a tunnel and later winds itself through it, forming a knot. It is interesting that the topology of the interfaces in the three-dimensional Ising model have

applications in quantum gravity⁹⁷.) In other systems, the statistics of all of these properties have been shown to be universal on long scales of length and time.

Continuing challenges

Despite recent developments, our understanding of crackling noise is far from complete. There are only a few systems^{28,32,48,49,53,54,78–80} where the renormalization-group framework has substantially explained even the behaviour of the numerical models. There are several other approaches^{63,64,87,98–100} that have been developed to study crackling noise, many of which share our view of developing effective descriptions on long scales of length and time. But the successes remain dwarfed by the bewildering variety of systems that crackle. Achieving a global perspective on the universality classes for crackling noise remains an open challenge.

An even more important challenge is to make quantitative comparison between the theoretical models and experimental systems. We believe that going beyond power laws will be crucial in this endeavour. The past focus on critical exponents has sometimes been frustrating: it is too easy to find power laws over limited scaling ranges¹⁰¹, and too easy to find models which roughly fit them. It also seems unfulfilling, summarizing a complex morphology into a single critical exponent. We believe that measuring a power law is almost never definitive by itself: a power law in conjunction with evidence that the morphology is scale invariant (for example, a scaling collapse) is crucial. By aggressively pursuing quantitative comparisons of other, richer measures of morphology such as the universal scaling functions, we will be better able both to discriminate among theories and to ensure that a measured power law corresponds to a genuine scaling behaviour.

Another challenge is to start thinking about the key ways that these complex spatiotemporal systems differ from the phase transitions we understand from equilibrium critical behaviour. (The renormalization-group tools developed by our predecessors are seductively illuminating, and it is always easy to focus where the light is good.) For example, in several of these systems there are collective, dynamical ‘memory’ effects^{15,102–105} that may even have practical applications¹⁰⁶. The quest for a scaling theory of crackling phenomena needs to be viewed as part of the larger process of understanding the dynamics of these nonlinear, non-equilibrium systems.

A final challenge is to make the study of crackling noise profitable. Less noise from candy wrappers^{11–13} in cinemas and theatres is not the most pressing of global concerns. Making money from fluctuations in stock prices is already big business^{58,59}. Predicting earthquakes over the short term probably will not be feasible using these approaches¹⁰⁷, but longer-term prediction of impending large earthquakes may be both possible⁸⁶ and useful (for example, for guiding local building codes). Understanding that the large-scale behaviour relies on only a few emergent material parameters (disorder and external field for our model of magnetism) will lead to the study of how these parameters depend on the microphysics. We might dream, for example, of learning eventually how to shift an active earthquake fault into a harmless, continuously sliding regime by adding lubricants to the fault gouge. In the meantime, crackling noise is basic research at its elegant, fundamental best. □

- Kadanoff, L. P. Scaling laws for Ising models near T_c . *Physics* **2**, 263–272 (1966).
- Wilson, K. G. Problems in physics with many scales of length. *Sci. Am.* **241**, 140–157 (1979).
- Pfeuty, P. & Toulouse, G. *Introduction to the Renormalization Group and to Critical Phenomena* (Wiley, London, 1977).
- Yeomans, J. M. *Statistical Mechanics of Phase Transitions* (Oxford Univ. Press, Oxford, 1992).
- Fisher, M. E. Renormalization group theory: its basis and formulation in statistical physics. *Rev. Mod. Phys.* **70**, 653–681 (1998).
- Martin, P. C., Siggia, E. D. & Rose, H. A. Statistical dynamics of classical systems. *Phys. Rev. A*, **423–437** (1973).
- De Dominicis, C. Dynamics as a substitute for replicas in systems with quenched random impurities. *Phys. Rev. B* **18**, 4913–4919 (1978).
- Sompolinsky, H. & Zippelius, A. Relaxational dynamics of the Edwards-Anderson model and the mean-field theory of spin-glasses. *Phys. Rev. B* **25**, 6860–6875 (1982).
- Zippelius, A. Critical-dynamics of spin-glasses. *Phys. Rev. B* **29**, 2717–2723 (1984).
- Gutenberg, B. & Richter, C. F. *Seismicity of the Earth and Associated Phenomena* (Princeton Univ.

- Press, Princeton, 1954).
11. Houle, P. A. & Sethna, J. P. Acoustic emission from crumpling paper. *Phys. Rev. E* **54**, 278–283 (1996).
12. Kramer, E. M. & Lobkovsky, A. E. Universal power law in the noise from a crumpled elastic sheet. *Phys. Rev. E* **53**, 1465–1469 (1996).
13. Glanz, J. No hope of silencing the phantom crinklers of the opera. *New York Times* 1 June 2000, A14 (2000).
14. Sethna, J. P. Hysteresis and avalanches <<http://www.lassp.cornell.edu/sethna/hysteresis/hysteresis.html>> (1996).
15. Sethna, J. P. *et al.* Hysteresis and hierarchies: dynamics of disorder-driven first-order phase transformations. *Phys. Rev. Lett.* **70**, 3347–3351 (1993).
16. Burridge, R. & Knopoff, L. Model and theoretical seismicity. *Bull. Seismol. Soc. Am.* **57**, 3411–3471 (1967).
17. Rice, J. R. & Ruina, A. L. Stability of steady frictional slipping. *J. Appl. Mech.* **50**, 343 (1983).
18. Carlson, J. M. & Langer, J. S. Mechanical model of an earthquake fault. *Phys. Rev. A* **40**, 6470–6484 (1989).
19. Bak, P. & Tang, C. Earthquakes as a self-organized critical phenomenon. *J. Geophys. Res.* **94**, 15635–15637 (1989).
20. Chen, K., Bak, P. & Obukhov, S. P. Self-organized criticality in a crack-propagation model of earthquakes. *Phys. Rev. A* **43**, 625–630 (1991).
21. Olami, Z., Feder, H. J. S. & Christensen, K. Self-organized criticality in a continuous, nonconservative cellular automaton modeling earthquakes. *Phys. Rev. Lett.* **68**, 1244–1247 (1992).
22. Miltenberger, P., Sornette, D. & Vanette, C. Fault self-organization and optimal random paths selected by critical spatiotemporal dynamics of earthquakes. *Phys. Rev. Lett.* **71**, 3604–3607 (1993).
23. Crowie, P. A., Vanette, C. & Sornette, D. Statistical physics model for the spatiotemporal evolution of faults. *J. Geophys. Res. Solid Earth* **98**, 21809–21821 (1993).
24. Carlson, J. M., Langer, J. S. & Shaw, B. E. Dynamics of earthquake faults. *Rev. Mod. Phys.* **66**, 657–670 (1994).
25. Myers, C. R., Shaw, B. E. & Langer, J. S. Slip complexity in a crustal-plane model of an earthquake fault. *Phys. Rev. Lett.* **77**, 972–975 (1996).
26. Shaw, B. E. & Rice, J. R. Existence of continuum complexity in the elastodynamics of repeated fault ruptures. *J. Geophys. Res.* **105**, 23791–23810 (2000).
27. Ben-Zion, Y. & Rice, J. R. Slip patterns and earthquake populations along different classes of faults in elastic solids. *J. Geophys. Res.* **100**, 12959–12983 (1995).
28. Fisher, D. S., Dahmen, K., Ramanathan, S. & Ben-Zion, Y. Statistics of earthquakes in simple models of heterogeneous faults. *Phys. Rev. Lett.* **78**, 4885–4888 (1997).
29. Fisher, D. S. Threshold behavior of charge-density waves pinned by impurities. *Phys. Rev. Lett.* **50**, 1486–1489 (1983).
30. Fisher, D. S. Sliding charge-density waves as a dynamic critical phenomenon. *Phys. Rev. B* **31**, 1396–1427 (1985).
31. Littlewood, P. B. Sliding charge-density waves: a numerical study. *Phys. Rev. B* **33**, 6694–6708 (1986).
32. Narayan, O. & Fisher, D. S. Critical behavior of sliding charge-density waves in 4- ϵ dimensions. *Phys. Rev. B* **46**, 11520–11549 (1992).
33. Middleton, A. A. & Fisher, D. S. Critical behavior of charge-density waves below threshold: numerical and scaling analysis. *Phys. Rev. B* **47**, 3530–3552 (1993).
34. Myers, C. R. & Sethna, J. P. Collective dynamics in a model of sliding charge-density waves. I. Critical behavior. *Phys. Rev. B* **47**, 11171–11192 (1993).
35. Thorne, R. E. Charge-density-wave conductors. *Phys. Today* **49**, 42–47 (1996).
36. Bak, P., Tang, C. & Wiesenfeld K. Self-organized criticality: an explanation for 1/f noise. *Phys. Rev. Lett.* **59**, 381–384 (1987).
37. Bak, P., Tang, C. & Wiesenfeld K. Self-organized criticality. *Phys. Rev. A* **38**, 364–374 (1988).
38. deGennes, P. G. *Superconductivity of Metals and Alloys* p. 83 (Benjamin, New York, 1966).
39. Feynman, R. P., Leighton, R. B. & Sands, M. *The Feynman Lectures on Physics* Vol. II Sect. 37–3 (Addison Wesley, Reading, MA, 1963–1965).
40. Jaeger, H. M., Liu, C. & Nagel, S. R. Relaxation at the angle of repose. *Phys. Rev. Lett.* **62**, 40–43 (1989).
41. Nagel, S. R. Instabilities in a sandpile. *Rev. Mod. Phys.* **64**, 321–325 (1992).
42. Tewari, S. *et al.* Statistics of shear-induced rearrangements in a two-dimensional model foam. *Phys. Rev. E* **60**, 4385–4396 (1999).
43. Solé, R. V. & Manrubia, S. C. Extinction and self-organized criticality in a model of large-scale evolution. *Phys. Rev. E* **54**, R42–R45 (1996).
44. Newman, M. E. J. Self-organized criticality, evolution, and the fossil extinction record. *Proc. R. Soc. Lond. B* **263**, 1605–1610 (1996).
45. Newman, M. E. J. & Palmer, R. G. Models of extinction: a review. Preprint [adap-org/9908002](http://xxx.lanl.gov) at <<http://xxx.lanl.gov>> (1999).
46. Cieplak, M. & Robbins, M. O. Dynamical transition in quasistatic fluid invasion in porous media. *Phys. Rev. Lett.* **60**, 2042–2045 (1988).
47. Koiller, B. & Robbins, M. O. Morphology transitions in three-dimensional domain growth with Gaussian random fields. *Phys. Rev. B* **62**, 5771–5778 (2000).
48. Nattermann, T., Stepanow, S., Tang, L. H. & Leschhorn N. Dynamics of interface depinning in a disordered medium. *J. Phys. II (Paris)* **2**, 1483–1488 (1992).
49. Narayan, O. & Fisher, D. S. Threshold critical dynamics of driven interfaces in random media. *Phys. Rev. B* **48**, 7030–7042 (1993).
50. Leschhorn, H., Nattermann, T., Stepanow, S. & Tang, L.-H. Driven interface depinning in a disordered medium. *Ann. Phys. (Leipzig)* **6**, 1–34 (1997).
51. Roters, L., Hucht, A., Lubeck, S., Nowak, U. & Usadel, K. D. Depinning transition and thermal fluctuations in the random-field Ising model. *Phys. Rev. E* **60**, 5202–5207 (1999).
52. Field, S., Witt, J., Nori, F. & Ling, X. Superconducting vortex avalanches. *Phys. Rev. Lett.* **74**, 1206–1209 (1995).
53. Ertas, D. & Kardar, M. Anisotropic scaling in depinning of a flux line. *Phys. Rev. Lett.* **73**, 1703–1706 (1994).
54. Ertas, D. & Kardar, M. Anisotropic scaling in threshold critical dynamics of driven directed lines. *Phys. Rev. B* **53**, 3520–3542 (1996).
55. Lilly, M. P., Wootters, A. H. & Hallock, R. B. Spatially extended avalanches in a hysteretic capillary condensation system: superfluid He-4 in nuclepore. *Phys. Rev. Lett.* **77**, 4222–4225 (1996).
56. Guyer, R. A. & McCall, K. R. Capillary condensation, invasion percolation, hysteresis, and discrete memory. *Phys. Rev. B* **54**, 18–21 (1996).
57. Ortin, J. *et al.* Experiments and models of avalanches in martensites. *J. Phys. IV (Paris)* **5**, 209–214 (1995).
58. Bouchaud, J. P. Power-laws in economy and finance: some ideas from physics. (Proc. Santa Fe Conf. Beyond Efficiency.) *J. Quant. Finance* (in the press); also available as preprint cond-mat/0008103 at <<http://xxx.lanl.gov>>.
59. Bak, P., Paczuski, M. & Shubik, M. Price variations in a stock market with many agents. *Physica A* **246**, 430–453 (1997).
60. Lu, E. T., Hamilton, R. J., McTiernan, J. M. & Bromond, K. R. Solar flares and avalanches in driven dissipative systems. *Astrophys. J.* **412**, 841–852 (1993).
61. Carreras, B. A., Newman, D. E., Dobson, I. & Poole, A. B. Initial evidence for self-organized criticality in electrical power system blackouts. In *Proc. 33rd Hawaii Int. Conf. Syst. Sci.* (ed. Sprague, R. H. Jr) (IEEE Comp. Soc., Los Alamitos, CA, 2000).
62. Sachtjen, M. L., Carreras, B. A. & Lynch, V. E. Disturbances in a power transmission system. *Phys. Rev. E* **61**, 4877–4882 (2000).
63. Carlson, J. M. & Doyle, J. Highly optimized tolerance: a mechanism for power laws in designed systems. *Phys. Rev. E* **60**, 1412–1427 (1999).
64. Carlson, J. M. & Doyle, J. Highly optimized tolerance: robustness and design in complex systems. *Phys. Rev. Lett.* **84**, 2529–2532 (2000).
65. Newman, M. The power of design. *Nature* **405**, 412–413 (2000).
66. Galam, S. Rational group decision making: a random field Ising model at T=0. *Physica A* **238**, 66–80 (1997).
67. Petri, A., Paparo, G., Vespignani, A., Alippi, A. & Costantini, M. Experimental evidence for critical dynamics in microfracturing processes. *Phys. Rev. Lett.* **73**, 3423–3426 (1994).
68. Garcimartin, A., Guarino, A., Bellon, L. & Ciliberto, S. Statistical properties of fracture precursors. *Phys. Rev. Lett.* **79**, 3202–3205 (1997).
69. Curtin, W. A. & Scher, H. Analytic model for scaling of breakdown. *Phys. Rev. Lett.* **67**, 2457–2460 (1991).
70. Herrman, H. J. & Roux, S. (eds) *Statistical Models for the Fracture of Disordered Media* (North Holland, Amsterdam, 1990).
71. Chakrabarti, B. K. & Benguigui, L. G. *Statistical Physics of Fracture and Breakdown in Disordered Systems* (Clarendon, Oxford, 1997).
72. Zapperi, S., Ray, P., Stanley, H. E. & Vespignani, A. First-order transition in the breakdown of disordered media. *Phys. Rev. Lett.* **78**, 1408–1411 (1997).
73. Perković, O., Dahmen, K. A. & Sethna, J. P. Avalanches, Barkhausen noise, and plain old criticality. *Phys. Rev. Lett.* **75**, 4528–4531 (1995).
74. Kuntz, M. C., Perković, O., Dahmen, K. A., Roberts, B. W. & Sethna, J. P. Hysteresis, avalanches, and noise: numerical methods. *Comput. Sci. Eng.* **1**, 73–81 (1999).
75. Kuntz, M. C. & Sethna, J. P. Hysteresis, avalanches, and noise: numerical methods <<http://www.lassp.cornell.edu/sethna/hysteresis/code/>> (1998).
76. Perković, O., Dahmen, K. A. & Sethna, J. P. Disorder-induced critical phenomena in hysteresis: numerical scaling in three and higher dimensions. *Phys. Rev. B* **59**, 6106–6119 (1999).
77. Berger, A., Inomata, A., Jiang, J. S., Pearson, J. E. & Bader, S. D. Experimental observation of disorder-driven hysteresis-loop criticality. *Phys. Rev. Lett.* **85**, 4176–4179 (2000).
78. Dahmen, K. A. & Sethna, J. P. Hysteresis, avalanches, and disorder induced critical scaling: a renormalization group approach. *Phys. Rev. B* **53**, 14872–14905 (1996).
79. da Silva, R. & Kardar, M. Critical hysteresis for N-component magnets. *Phys. Rev. E* **59**, 1355–1367 (1999).
80. Dahmen, K. A. & Sethna, J. P. Hysteresis loop critical exponents in 6- ϵ dimensions. *Phys. Rev. Lett.* **71**, 3222–3225 (1993).
81. Visscher, P. B. Renormalization-group derivation of Navier-Stokes equation. *J. Stat. Phys.* **38**, 989–1013 (1985).
82. Kadanoff, L. P., McNamara, G. R. & Zanetti, G. From automata to fluid flow: comparisons of simulation and theory. *Phys. Rev. A* **40**, 4527–4541 (1989).
83. Hwa, T. & Kardar, M. Dissipative transport in open systems: an investigation of self-organized criticality. *Phys. Rev. Lett.* **62**, 1813–1816 (1989).
84. Grinstein, G., Lee, D.-H. & Sachdev, S. Conservation laws, anisotropy, and “self-organized criticality” in noisy non-equilibrium systems. *Phys. Rev. Lett.* **64**, 1927–1930 (1990).
85. Sornette, D. Sweeping of an instability—an alternative to self-organized criticality to get power laws without parameter tuning. *J. Phys. I (Paris)* **4**, 209–221 (1994).
86. Sykes, L. R., Shaw, B. E. & Scholz, C. H. Rethinking earthquake prediction. *Pure Appl. Geophys.* **155**, 207 (1999).
87. Carlson, J. M., Chayes, J. T., Grannan, E. R. & Swindle, G. H. Self-organized criticality and singular diffusion. *Phys. Rev. Lett.* **65**, 2547–2550 (1990).
88. Urbach, J. S., Madison, R. C. & Markert, J. T. Interface depinning, self-organized criticality, and the Barkhausen effect. *Phys. Rev. Lett.* **75**, 276–279 (1995).
89. Narayan, O. Self-similar Barkhausen noise in magnetic domain wall motion. *Phys. Rev. Lett.* **77**, 3855–3857 (1996).
90. Zapperi, P., Cizeau, P., Durin, G. & Stanley, H. E. Dynamics of a ferromagnetic domain wall: avalanches, depinning transition, and the Barkhausen effect. *Phys. Rev. B* **58**, 6353–6366 (1998).
91. Pazmandi F., Zarand G. & Zimanyi G. T. Self-organized criticality in the hysteresis of the Sherrington-Kirkpatrick model. *Phys. Rev. Lett.* **83**, 1034–1037 (1999).
92. Pazmandi F., Zarand G. & Zimanyi G. T. Self-organized criticality in the hysteresis of the Sherrington-Kirkpatrick model. *Physica B* **275**, 207–211 (2000).
93. Perković, O., Dahmen, K. A. & Sethna, J. P. Disorder-induced critical phenomena in hysteresis: a numerical scaling analysis. Preprint cond-mat/9609072, appendix A, at <<http://xxx.lanl.gov>> (1996).
94. Kuntz, M. C. & Sethna, J. P. Noise in disordered systems: the power spectrum and dynamic exponents in avalanche models. *Phys. Rev. B* **62**, 11699–11708 (2000).
95. Spasojević, D., Bukvić, S., Milošević, S. & Stanley, H. E. Barkhausen noise: elementary signals, power laws, and scaling relations. *Phys. Rev. E* **54**, 2531–2546 (1996).
96. Family, F., Vicsek, T. & Meakin, P. Are random fractal clusters isotropic? *Phys. Rev. Lett.* **55**, 641–644 (1985).
97. Dotsenko, V. S. *et al.* Critical and topological properties of cluster boundaries in the 3D Ising model. *Phys. Rev. Lett.* **71**, 811–814 (1993).
98. Kadanoff, L. P., Nagel, S. R., Wu, L. & Zhou, S.-M. Scaling and universality in avalanches. *Phys. Rev. A* **39**, 6524–6537 (1989).
99. Dhar, D. The Abelian sandpile and related models. *Physica A* **263**, 4–25 (1999).
100. Paczuski, M., Maslov, S. & Bak, P. Avalanche dynamics in evolution, growth, and depinning models. *Phys. Rev. E* **414**–443 (1996).
101. Malcai, O., Lidar, D. A., Biham, O. & Avnir, D. Scaling range and cutoffs in empirical fractals. *Phys. Rev. E* **56**, 2817–2828 (1997).
102. Fleming, R. M. & Schneemeyer, L. F. Observation of a pulse-duration memory effect in $K_{0.30}MoO_3$. *Phys. Rev. Lett.* **33**, 2930–2932 (1986).

103. Coppersmith, S. N. & Littlewood, P. B. Pulse-duration memory effect and deformable charge-density waves. *Phys. Rev. B* **36**, 311–317 (1987).
104. Middleton, A. A. Asymptotic uniqueness of the sliding state for charge-density waves. *Phys. Rev. Lett.* **68**, 670–673 (1992).
105. Amengual, A. *et al.* Systematic study of the martensitic transformation in a Cu–Zn–Al alloy—reversibility versus irreversibility via acoustic emission. *Thermochim. Acta* **116**, 195–308 (1987).
106. Perković, O. & Sethna, J. P. Improved magnetic information storage using return-point memory. *J. Appl. Phys.* **81**, 1590–1597 (1997).
107. Pepke, S. L., Carlson, J. M. & Shaw, B. E. Prediction of large events on a dynamical model of a fault. *J. Geophys. Res.* **99**, 6769 (1994).
108. Council of the National Seismic System. Composite Earthquake Catalog Archive <<http://www.cnss.org>> (2000).
109. US Geological Survey National Earthquake Information Center. Earthquake information for the world <<http://www.neic.cr.usgs.gov>> (2001).
110. Sethna, J. P., Kuntz, M. C., & Houle, P. A. Crackling noise <<http://simscience.org/crackling>>. (1999).
111. Brézin E. & De Dominicis C. Dynamics versus replicas in the random field Ising model. *C.R. Acad. Sci. II* **327**, 383–390 (1999).
112. Cote, P. J. & Meisel, L. V. Self-organized criticality and the Barkhausen effect. *Phys. Rev. Lett.* **67**, 1334–1337 (1991).
113. Meisel, L. V. & Cote, P. J. Power laws, flicker noise, and the Barkhausen effect. *Phys. Rev. B* **46**, 10822–10828 (1992).
114. Stierstadt, K. & Boeckh, W. Die Temperaturabhängigkeit des Magnetischen Barkhauseneffekts. 3. Die Sprunggrößenverteilung längs der Magnetisierungskurve. *Z. Phys.* **186**, 154 (1965).
115. Bertotti, G., Durin, G. & Magni, A. Scaling aspects of domain wall dynamics and Barkhausen effect in ferromagnetic materials. *J. Appl. Phys.* **75**, 5490–5492 (1994).
116. Bertotti, G., Fiorillo, F. & Montorsi, A. The role of grain size in the magnetization process of soft magnetic materials. *J. Appl. Phys.* **67**, 5574–5576 (1990).
117. Lieneweg, U. Barkhausen noise of 3% Si-Fe strips after plastic deformation. *IEEE Trans. Magn.* **10**, 118–120 (1974).
118. Lieneweg, U. & Grosse-Nobis, W. Distribution of size and duration of Barkhausen pulses and energy spectrum of Barkhausen noise investigated on 81% nickel-iron after heat treatment. *Int. J. Magn.* **3**, 11–16 (1972).
119. Bittel, H. Noise of ferromagnetic materials. *IEEE Trans. Magn.* **5**, 359–365 (1969).
120. Montalenti, G. Barkhausen noise in ferromagnetic materials. *Z. Angew. Phys.* **28**, 295–300 (1970).
121. Durin, G. & Zapperi, S. Scaling exponents for Barkhausen avalanches in polycrystalline and amorphous ferromagnets. *Phys. Rev. Lett.* **84**, 4705–4708 (2000).
122. Petta, J. R. & Weissmann, M. B. Barkhausen pulse structure in an amorphous ferromagnet: characterization by high-order spectra. *Phys. Rev. E* **57**, 6363–6369 (1998).
123. Alessandro, B., Beatrice, C., Bertotti, G., & Montorsi, A. Domain-wall dynamics and Barkhausen effect in metallic ferromagnetic materials. 1. Theory. *J. Appl. Phys.* **68**, 2901–2907 (1990).
124. Alessandro, B., Beatrice, C., Bertotti, G. & Montorsi, A. Domain-wall dynamics and Barkhausen effect in metallic ferromagnetic materials. 2. Experiment. *J. Appl. Phys.* **68**, 2908–2915 (1990).
125. Walsh, B., Austvold, S. & Proksch, R. Magnetic force microscopy of avalanche dynamics in magnetic media. *J. Appl. Phys.* **84**, 5709–5714 (1998).
126. Kryscak, L. C. & Maynard, J. D. Evidence for the role of propagating stress waves during fracture. *Phys. Rev. Lett.* **81**, 4428–4431 (1998).

Acknowledgements

The perspective on this field described in this paper grew out of a collaboration with M. Kuntz. We thank A. Mehta for supplying the data for Fig. 7b, and D. Dolgert, M. Newman, J.-P. Bouchaud, L. C. Kryscak, D. Fisher and J. Thorpe for helpful comments and references. This work was supported by NSF grants, the Cornell Theory Center and IBM.

Noise to order

Troy Shinbrot & Fernando J. Muzzio

Department of Chemical & Biochemical Engineering, Rutgers University, Piscataway, New Jersey 08854, USA (e-mail: shinbrot@sol.rutgers.edu)

Patterns in natural systems abound, from the stripes on a zebra to ripples in a riverbed. In many of these systems, the appearance of an ordered state is not unexpected as the outcome of an underlying ordered process. Thus crystal growth, honeycomb manufacture and floret evolution generate regular and predictable patterns. Intrinsically noisy and disordered processes such as thermal fluctuations or mechanically randomized scattering generate surprisingly similar patterns. Here we discuss some of the underlying mechanisms believed to be at the heart of these similarities.

If one rubs a piece of flannel across a blackboard, chalk from the board becomes distributed uniformly across both flannel and board, erasing a pattern drawn earlier. The converse does not occur: rubbing flannel on a blackboard does not cause chalk to leave a more ordered state than found initially. This is a crude but straightforward consequence of the second law of thermodynamics, which roughly stated dictates that disorder tends to increase. On the other hand — as is well known to any parent who has organized a child's party — if one rubs the same flannel on a plastic balloon, charge transfer can take place between cloth and balloon, against the apparent electrostatic gradient¹, leaving the balloon highly charged and predisposed to be stuck to a nearby wall. How is it that chalk, when rubbed, tends to become more uniformly distributed, whereas plastic, when rubbed, causes surface charges to become more strongly separated?

This example, although unsophisticated, is representative of a growing number of problems in modern physics in which noisy processes (for example, rubbing) can result in increased ordering². Noise-induced ordering can take many forms, ranging from mere separation of different materials into a heterogeneous final state (as in the example of charges on a balloon) to the formation of a rich variety of regular patterns, including stripes, squares, hexagons and spirals. These phenomena are seen in catalysis³, cosmology^{4–6}, marine biology⁷, reactive mixing^{8–10}, colloidal chemistry¹¹ and geophysics^{12–14}. In some of these problems, most notably in colloids research^{11,15}, mechanisms such as steric stabilization have long been understood, whereas in others the causes and implications of spontaneous pattern formation continue to generate developments that are theoretically meaningful and of practical importance. In this article we review some of the high points in historical progress in the field and focus on research in two of the fastest growing areas: granular physics and phase separation.

The use of noise for productive purposes has a long and colourful history. For over a century it has been reported by sailors that disordered raindrops falling on the ocean will calm rough seas^{16–18}. By comparison, noise in the form of heat has for many decades been used in the electronics industry to instigate charge release from surfaces and so amplify signals (for example, in vacuum tubes), deposit ions (for example, in thin-film fabrication), or generate displays (for example, in cathode ray tubes). Noise in the form of tiny disturbances provokes instabilities that lead to regular oscillations in systems ranging from musical instruments and humming power lines to road corrugation and catastrophic structural collapse¹⁹. Noise is also important in

increasing signal-processing effectiveness in applications as diverse as neuroprocessing and palaeoclimatology²⁰, and the addition of noise to pattern-forming systems can actually enhance the pattern-formation mechanism in several problems^{10,21,22}. In other applications, aspects of noise can improve the regularity of long-term biological rhythms²³.

For our purposes, it is useful to differentiate between problems in which noise enhances or initiates an existing process, and qualitatively different problems in which order would not appear at all without the presence of noise. This distinction is not always clear-cut, as in the case of convection rolls that are arguably both initiated and sustained by noise in the form of heat. Nevertheless, the issues we seek to explore are how noise itself produces ordered states, and what can be learnt about patterns in natural systems from the mechanisms that permit this to occur.

Noise, order and the second law of thermodynamics

Noise at the macroscopic scale, or equivalently diffusion at the microscopic scale, has long been investigated in chemical and thermodynamic systems. Historically, study of the emergence of periodic patterns from diffusive processes has had a lively development, beginning arguably with the travails of Boris Belousov, who in the 1950s reported to a disbelieving scientific community²⁴ that chemical reactions can evolve periodically in time towards a heterogeneous final state (see Box 1). The reason for conventional disbelief in this, now well accepted, result lies at the heart of the present discussion. That is, diffusion is characterized by a monotonic tendency towards homogeneity as dictated by the equation:

$$\frac{\partial C_A}{\partial t} = D_A \nabla^2 C_A \quad (1)$$

where C_A is the concentration of a substance of interest, and D_A is a diffusion coefficient. It is self-evident that this equation is first order in time and prescribes a monotonic change in C_A . It therefore seems paradoxical that diffusion combined with any sequence of terminal reactions can produce evolving patterns. Heuristic analysis on the microscale produces a similar result: diffusion is nothing more than random motion of particles caused by brownian agitation, and it is not at all obvious that anything other than homogeneity can result from such motion. Indeed, the second law of thermodynamics hinges on this very principle. We stress at the outset that the second law is robust in the face of all of the examples that we will present, as can be confirmed through a careful accounting of energy and entropy²⁵. Nevertheless, the overt behaviours found in these problems can be strikingly counterintuitive.

As an example, consider a macroscale version of the Maxwell's Demon problem. Maxwell's Demon is the

paradigmatic thought experiment used to challenge the second law²⁶, and consists of the hypothetical process of extracting heat from a gas by opening an imaginary door separating two chambers only when high-speed (that is, hot) molecules approach the door from, say, the right. In this way, heat could conceivably be transferred from one chamber to its neighbour without the expenditure of energy. A meticulous analysis of the actual energy required to implement such a machine reveals that the second law is sound because the minimum expenditure of energy needed to detect and respond to fast particles is

Box 1

Liesegang bands

Geology is rich with patterned states, ranging from large sand-dune patterns⁷⁶ and cellular stone formations² to striated rocks⁷⁷, gems^{15,78} and sedimentary structures^{79,80}. Many of these patterns are so complex that scientists continue to debate whether they were produced by organized, biological processes or disorganized, inorganic means^{12,81–83}. One of the earliest situations where geological patterns governed by inorganic diffusion have been analysed is in 'Liesegang' bands, proposed to be responsible for the striped appearance of iris agates⁷⁸. These bands, shown in the figure below, occur in gels or other media where convection is suppressed and only diffusive mixing is possible³⁵. In its simplest analysis, the mechanism is as follows. A solute (added from above in the experiment illustrated) diffuses through a static medium that may contain a second reactant. If the solute or its reaction product is present in sufficient concentration, it can supersaturate the medium and subsequently precipitate out of solution in a local region. The precipitate depletes the region of solute, thus delaying its accumulation downstream. The solute diffuses through the obstacle presented by the precipitate to a distance downstream where it once again accumulates to reach supersaturation levels. At this point the cycle repeats itself. In the wake of these cycles of diffusion, supersaturation and precipitation, bands of precipitate and low-concentration solute are left behind.

This analysis can be embellished by including the effects of unavoidable complications such as differential solubilities and crystallization speeds as well as weak convective cycles^{79,80}. Notwithstanding these, sometimes significant, refinements, a precipitation–diffusion cycle seems to be the dominant mechanism at work in the formation of Liesegang patterns. We note in particular that, as presaged by Belousov (see text), although the kinetics are governed by a first-order diffusion equation, spatial pattern formation occurs as a result of coupling between this first-order equation and a more complicated, supersaturation–precipitation cycle.



Box 1 Figure Liesegang patterns in a cylinder containing agar gel with dissolved K_2CrO_4 to which $CuSO_4$ has been added from above. See <http://qsad.bu.edu/ogaf/lies/> for details of this system suitable for instructional use. (Photo: P. Garik, M. B. Gillespie and K. Brecher, Mathematics Education Center, Boston University.)

nonzero and in fact exactly equals that predicted by quantitative forms of the law²⁵.

But in the case of granular physics, where much recent research on noise and pattern formation has focused, a surprising result is obtained. This result, first discussed²⁷ in 1996 and since analysed in detail²⁸, is duplicated in Fig. 1, where we display the results of an experiment in which a number of steel beads were initially distributed uniformly across a vibrated acrylic container. The container is separated into two identical chambers by a foamboard barrier containing an open window near its bottom.

Despite being distributed uniformly at the start, after a short time (2–3 minutes) most of the beads migrate spontaneously to one chamber or the other. This separates the system into a first chamber containing a sparse and gas-like state of high-speed particles, and a second containing a dense and close-packed state of nearly motionless beads. This separation of 'hot' and 'cold' states occurs in the presence of noisy agitation in the form of a macroscopic analogue of brownian motion: experiments²⁹ and simulations^{30,31} have shown that the particles' horizontal velocities are effectively randomized when vibrated energetically in this way.

Unlike the classic Maxwell's Demon experiment, the window remains open, the two sides are identical, and particles are free to travel between the chambers at all times. Also unlike the classic case — and central to the ordering mechanism in this and all other problems — this system has an intrinsic and pronounced source of dissipation. The dissipation here results from inelastic collisions between the beads, and the mechanism by which this dissipation leads to spontaneous heating and freezing of adjacent chambers is straightforward: each time a bead suffers a collision, it loses kinetic energy to internal heat owing to inelasticity of the bead material. Because energy is lost with every collision, energy delivered to the bottom of a bed of beads must be lost exponentially with height (as measured in numbers of beads) through the bed, and a small surplus of beads in one of the two chambers can produce a large excess in dissipation in that chamber. This means that once, by chance, a small surplus of beads appears on either side of the barrier, that side will lose energy more rapidly than the opposing side, and will therefore be able to eject fewer beads through the window than it receives. For suitable vibration conditions²⁸, this instability accelerates rapidly, causing one chamber to lose so much energy that its beads effectively become frozen, while the other chamber loses increasingly more particles, and each particle remaining experiences fewer collisions and so remains highly energetic.

This is the state shown in Fig. 1, where a 'hot gas' has apparently separated spontaneously from a 'frozen solid'. This separation occurs only in this macroscopic view: a full analysis both of the actual heating of the grains and of the heating of the environment as a result of energy input through an external shaker would readily reveal that entropy overall in this nonequilibrium system increases (markedly) with time. Nevertheless, the apparent effect is that a disordered and homogenous initial state has been transformed into an ordered and heterogeneous state as a result of noisy agitation.

Patterns from microscale diffusion

Against expectation, ordering of macroscopic particles into gas-like and frozen states can arise spontaneously. The ingredients required to produce this outcome are a source of noisy energy (here the shaking of the container) and a source of dissipation (the inelasticity of the particles). The same ingredients on the microscopic scale produce regular ordering in a variety of natural situations as well. Historically, one of the earliest mentions of this kind of ordering — and one of the most elegant systems in which it can be produced — is in the production of layered geological specimens, where 'Liesegang bands' of precipitating reactants can form spontaneously, resulting in stripes and other patterns in otherwise static gels (see Box 1).

The possibility that patterns could form in natural systems due only to reaction combined with molecular noise (that is, diffusion) was proposed in 1952 by Alan Turing. The emergence of these

Figure 1 Separation of hot and cold beads in a granular 'Maxwell's Demon' experiment.

a, Snapshot of spontaneous separation between gas-like and solid-like steel beads (diameter 2 mm) from an experiment in which an acrylic container is vibrated sinusoidally at 10 Hz and a maximum acceleration of 1.3g. Initially the beads are distributed uniformly

across the container, but owing to collisional inelasticity they migrate to whichever side randomly acquires a slight excess of beads. **b**, Schematic of container which is divided into two chambers by a barrier containing a window 6 mm high, beginning 6 mm from the bottom of the container.



patterns from an initially homogeneous state seems to run counter to the simple analysis of equation (1). The paradoxical appearance of patterns in experiments occurs when two reactants diffuse at different rates, thereby generating differing reaction outcomes as their gradients change. This is a subtle effect with profound consequences, and permits the transmission of neuronal waves in biological systems such as the heart^{32,33}, as well as the emergence of complex ecological³⁴, chemical^{8–10} and bacterial² patterns. In all of these varied problems, equation (1) is modified first by the inclusion of a reacting term and second by a coupling to a second 'reaction-diffusion' equation describing another species. In neurons, the two equations relate to discharge and repolarization across membranes; in ecology they relate to differential flow and uptake of nutrients; and in catalytic and other reactions they relate to different diffusivities of constituent reactants and by-products.

Example

One of the simplest scenarios in which this effect is seen is the following set of reactions, simulated^{9,35} in 1993 for reactants A and B:



Reaction such as (2a) are seen in biological problems such as when a nutrient, A, is added to a population of cells, B, yielding more cells, or in crystallization problems where an antisolvent, A, is added to a saturated solution to increase the abundance of a crystalline product, B.

Reaction (2b) represents a decay such as occurs during ageing or oxidation of biological or chemical systems.

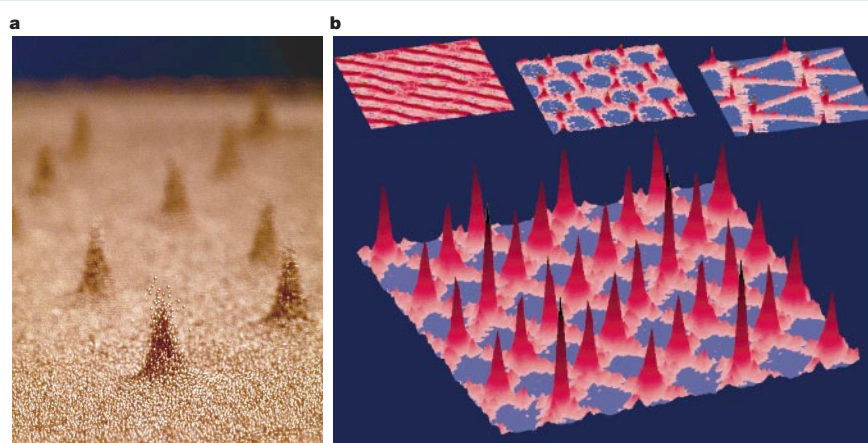
When reactant A is fed continuously and uniformly into the system, these reactions, suitably normalized, can be described by the coupled equations:

$$\frac{\partial C_A}{\partial t} = D_A \nabla^2 C_A - C_A C_B^2 + S_A (1 - C_A) \quad (3a)$$

$$\frac{\partial C_B}{\partial t} = D_B \nabla^2 C_B + C_A C_B^2 - (S_A + S_B) C_B \quad (3b)$$

where $S_{A,B}$ are constants regulating feed and reaction rates and $D_{A,B}$ are diffusion coefficients. The first term on the right side of each equation describes diffusion of each reactant, the second describes the reaction (2a), and the third describes feed of reactant A and consumption of A and B through reactions (2a) and (2b). When $D_A > D_B$, this system can give rise to complex patterns including stripes, spots, labyrinths and spirals. In this system, the heuristic mechanism for pattern formation is (similar to that of Liesegang bands discussed in Box 1) that reactant A spreads more rapidly than reactant B, and when the feed rate defined by S_A is sufficiently small, A becomes depleted in local regions through reaction (2a). When this occurs, the resulting excess of reactant B leads to a new product C in that region, as indicated in reaction (2b). This in turn produces a local depletion of B, enabling the feed to produce a local surplus of

Figure 2 Granular patterns from experiment and model. **a**, Array of 'oscillons' from a granular vibration experiment. (Photo: P. Umbanhowar and H. L. Swinney, Univ. Texas at Austin.) **b**, Patterns from simulations of 32,767 idealized particles that are randomized periodically and collide inelastically (see text). Plotted are densities of particles as a function of position, in square state (main plot: $f=0.5$, $V=2,500$, $T=1.25 \times 10^{-4}$), in stripes (left inset: $f=0.5$, $V=1,500$, $T=1.5 \times 10^{-4}$), in hexagons (centre inset: $f=0.5$, $V=2,000$, $T=2 \times 10^{-4}$) and in triangles (right inset: $f=0.5$, $V=3,500$, $T=2 \times 10^{-4}$).



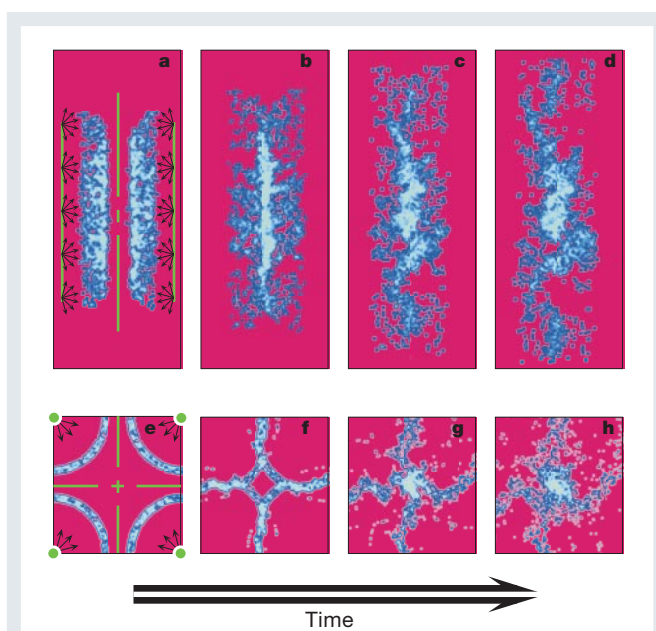


Figure 3 Stability of striped and square lattice patterns. Response of particles distributed near parallel stripes (a–d) or square gridpoints (e–h) to an instantaneous noisy impulse in the model simulation described in the text. The dashed green lines identify reflection symmetries of each configuration, and the solid green lines and small black arrows indicate locations and directions of initial particles. a–b, The striped initial state generates a new stripe at symmetry line. c–d, The new stripe is unstable to transverse fluctuations, causing it to decompose into clusters. e–h, The instability of the striped state causes clusters to form at the intersections of symmetry lines, thus stabilizing the square and other patterned states.

unreacted A. This surplus diffuses towards regions that have become depleted of A, and the cycle repeats itself.

Although this scenario might appear to be implausibly complicated at first blush, with some reflection the underlying mechanism can be seen to be so simple as to actually be quite probable. This is perhaps most clear in problems involving excitable media. In these media, a biological or chemical fuel is consumed in a local region, and the region cannot support further activity until the fuel has been replenished. Thus as an activation wave travels through neuronal tissue, cells left in its wake must wait for a refractory period to elapse before being able to fire again³². Sequential wave fronts, separated characteristically by the product of the wave speed and the refractory period, can thus be supported in such media. A similar explanation is obtained in combustion^{36,37} or catalysis³ experiments, when fuel on an extended surface is consumed more rapidly than the available oxygen can sustain. In this case, the reaction becomes extinguished by its own by-products, which must diffuse away before new reactions can proceed.

In reaction-diffusion problems, the difference between diffusivities D_A and D_B is central to the emergence of patterns. In a system with identical diffusivities, the spatial term can be eliminated from equations (3a) and (3b), leaving a single first-order equation that is incapable of describing spatial patterns. But the case $D_A \equiv D_B$ is non-generic and so there are numerous practical problems that are subject to pattern formation. The case $D_A < D_B$, on the other hand, is much more common, and in this situation, reactant A diffuses slower than heterogeneities can develop, thus stabilizing the system and leading to eventual homogeneity.

The reason for the required asymmetry of D_A and D_B is that B reacts nonlinearly, whereas A does not. This nonlinearity is a manifestation of the reproduction of B in reaction (2a), which augments local concentrations of B in much the same way that elastic dissipation promotes the concentrated frozen state in the granular Maxwell's Demon experiment. The same is not true of reactant A, which responds linearly and can increase its concentration only in

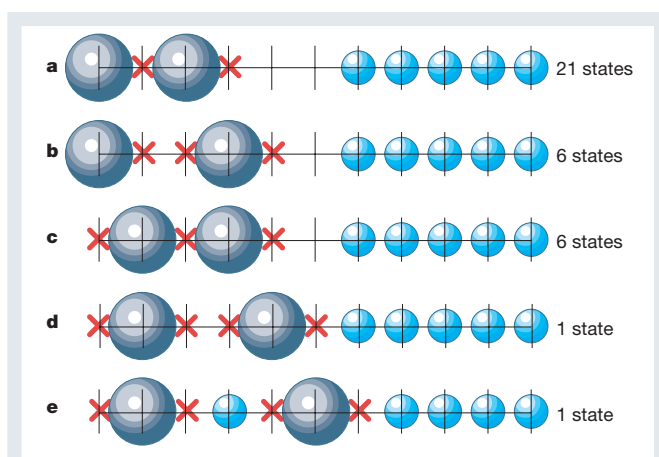


Figure 4 Entropic ordering example in simple one-dimensional lattice. Several of the 270 possible configurations consisting of 2 large (grey) and 5 small (blue) particles constrained to lie on a one-dimensional lattice of gridpoints. Excluded volumes surrounding the large particles are indicated by red x's, and the numbers of distinct states of small particles for each configuration of large particles depicted are indicated on the right. For randomly placed particles, states in which like particles lie together are favoured, and states in which large particles lie near an edge are greatly favoured.

response to the external supply. It is competition between diffusive gradients (that is, spreading caused by noise) and localization (resulting from nonlinear reactions) that produces patterns in reaction-diffusion systems. In the next section we describe larger-scale analogues of this same competition.

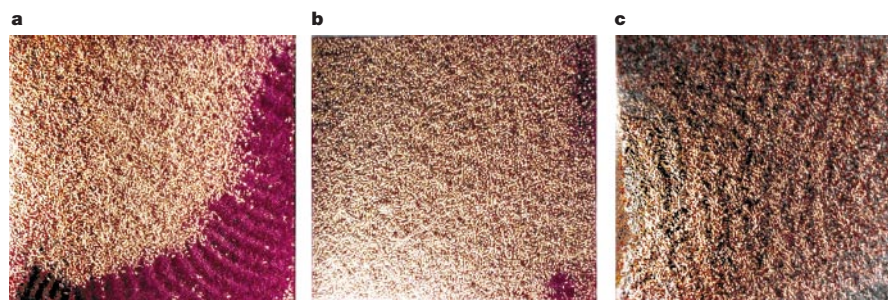
Patterns from macroscale noise

Patterns are seen on all scales, from the microscopic to the astrophysical. Thus complementing chemical and microbial² patterns, the regular spiral structure of many galaxies is believed to be due to dissipative collisions imposed on a background of initially random star locations^{5,6,38}. At larger scales still, scientific debate has intensified during the past decade over the possible presence of periodic patterns of clusters and superclusters of galaxies extending over as much as hundreds of millions of parsecs^{4,39,40}. These patterns are strongly reminiscent of results from simulations⁴¹ of inelastic hard spheres used to model granular flows, which we turn to next.

On laboratory scales, a variety of striking patterns have been reported^{42,43} in vibrated granular experiments similar to the granular Maxwell's Demon system of Fig. 1. Among the most intriguing are 'oscillons', which consist of a state of peaks and dips that alternate on successive vibration cycles⁴⁴. An array of oscillon peaks from a granular vibration experiment⁴⁴ is shown in Fig. 2a. As with chemical and astrophysical problems, many of the patterns seen in vibrated grains can be duplicated by models^{30,45–47} that combine the effects of noise or diffusion with dissipative or nonlinear interactions. Results of one such model are shown in Fig. 2b. Note that the square pattern in Fig. 2b differs from the oscillon array in Fig. 2a in that the peaks in this simulation appear in periodic lattices, whereas oscillons have been generated in both periodic and aperiodic arrangements^{44,46}.

Vibrated granular beds represent a particularly illuminating archetype for understanding the mechanism of pattern formation from noise. We have seen (Fig. 1) that the significance of dissipation for the generation of localized states is easily understood in granular beds. But localization (referred to as 'clustering' in the granular literature) is only part of the process: a clustered state does not necessarily display any regular pattern, such as the squares, stripes, hexagons or triangles shown in Fig. 2b. The other characteristic of importance is the regularity of the pattern itself. This regularity requires an additional ingredient, namely an underlying set of symmetries. Thus each of the patterns shown in Fig. 2b represents a particular symmetry. In

Figure 5 Possible transition from segregation to mixing in a vibrated granular bed. Possible granular transition between mixing and segregation of 1.8-mm (gold) and 0.8-mm (maroon) glass beads. Vibration is approximately sinusoidal with peak acceleration of 3.9g and frequency of 20 Hz. Under fixed conditions, the concentrations (as a weight fraction of the total granular bed) of small beads in the three snapshots are: **a**, 30%; **b**, 12%; **c**, 11%. Striped patterns, characteristic of vibrated grains, are evident in the figure but have no known role in the entropic segregation mechanism.



many problems, the source of these symmetries can be abstruse; in the granular case it is comparatively transparent.

Example

The model of Fig. 2b was engineered to capture the essential causes of pattern formation due to noise combined with dissipation. The model is a simplification of these two effects applied to the particular problem of vibrated beds of grains. High-speed photographs⁴⁸ of granular patterns reveal two distinct stages of motion. First, the supporting plate accelerates upwards to strike the granular bed. Patterns are at their sharpest at the moment of impact, and scatter apart thereafter. Second, the supporting plate accelerates downwards faster than gravity (accelerations of 3–8g are typical). The granular bed then separates from the supporting plate, and particles collide with one another to reform a patterned state, typically with the vertices of the new pattern at the centres of the previous pattern⁴⁹. Thus the simplified model for the system subjects a large number of identical discrete particles to two sequential processes: first, once per driving period, T , particles are scattered apart; and second, they suffer inelastic collisions thereafter³⁰.

The scattering process is simulated by instantaneously randomizing all particle velocities once per cycle. Randomization can be due to many different effects in problems as diverse as chemical waves and interstellar interactions; the randomization in the model of Fig. 2b is associated with multiple collisions between convex particles which effectively disorder particle vectors^{50,51}, in agreement with experimental and direct numerical data for the vibrated bed^{29,42}. Explicitly, particle velocities in Fig. 2b are reassigned periodically according to

$$\mathbf{U}_i \rightarrow f\mathbf{U}_i + V\eta_i \quad (4)$$

where $\mathbf{U}_i$ is the velocity of the i -th particle, η_i is a noise vector that is random in direction for each particle and is fixed in magnitude, and the arrow symbolizes evolution of the velocity after impact. The terms f and V are constants that respectively determine the particles' memories of their prior velocities and the strength of the impact. In the square pattern shown in Fig. 2b, we choose $f = 0.5$, $V = 2,500$, and we randomize velocities of each particle every 1.25×10^{-4} natural units (parameters for the other patterns shown are defined in the figure caption).

The dissipative process is intended to mimic the results of inelastic particle collisions, shown in Fig. 1 to generate spontaneous concentrations of particles^{41,49,52}. This too can be modelled in many ways; in Fig. 2b we show the results of a mean-field approach in which collisions are non-binary and follow the rule

$$\mathbf{U}_i \rightarrow \frac{(1 + n_i)\mathbf{U}_i + (1 - \varepsilon)n_i\mathbf{U}_j}{1 + n_i} \quad (5)$$

where $\mathbf{U}_i$ and $\mathbf{U}_j$ are respectively the velocities of the i -th particle and the mean velocity of its n_i neighbours within a specified radius, and ε

is the effective coefficient of restitution for collisions. When not under the influence of periodic randomization (equation (4)) or collisions with neighbours (equation (5)), particles travel ballistically in the plane. The model is strictly two-dimensional and may be thought of as a planar projection of the true three-dimensional problem. In Fig. 2b, we use $\varepsilon = 0.25$, and integrate trajectories of 32,767 initially randomly placed particles with a time step of 5×10^{-6} natural units. The figure shows the density of particles on a planar computational domain with periodic boundaries.

In addition to the square pattern, this simple model produces (insets to Fig. 2b) stripes, hexagons, triangles and other states as parameters such as the randomization period T and the amplitude V are varied. It may seem surprising that such a simplified model, containing little more than frequent randomization combined with dissipation, and entirely neglecting gravity, friction and other realistic effects, can produce an apparent wealth of spontaneously organized and highly structured patterns. As we discuss in the next section, an alternative view is that these patterns actually form *because* the system is randomized and dissipative. These features can by themselves dictate the formation of patterns independent of specifics of the problem under study. This is an extremely powerful conclusion that goes some way to explaining how so many different physical, chemical and biological systems develop such similar patterned states².

Role of symmetries

In the granular case, the mechanism for the formation and selection³⁰ of patterns, rather than disorganized clusters⁴¹, in the presence of noise is especially straightforward to analyse. Consider the case depicted in the top panels of Fig. 3, where a large number (4,000) of stationary particles, initially distributed near two parallel stripes, are instantaneously given a noise impulse of constant amplitude V and random direction. The random distribution of directions assures that particles will expand outwards from their initial configuration, and the half of particles that travel towards the centre line separating the initial stripes are plotted in Fig. 3a–d. These particles cannot avoid colliding, and when they do so they spawn a new stripe midway between the original ones (Fig. 3b). If the particle velocities are again randomized at close to this time (approximately D/V where D is the separation between the initial stripes), the new stripe will expand outwards as before. In a plane tiled initially with stripes, the striped state will reinforce itself, generating new stripes after each randomizing event, interlaced with the stripes before the event. This is what is seen both in experiments⁴⁸ and in simulations^{30,42}.

We emphasize that the only function of the noisy impulse in the pattern-formation mechanism is to cause high-concentration regions to expand outwards. From this perspective, any process that produces such an outwards expansion is equivalent — whether due to an instantaneous impulse, or to steady diffusion, or to some other (for example, higher-order) mechanism. It is for this reason that patterns caused by noise so closely resemble patterns found in deterministic systems and that patterns generated by different processes are so similar.

Which particular patterned state will predominate at given parameter values, on the other hand, is a question of stability, and here the specifics of a given problem enter into consideration. In the model system of Fig. 3, for example, the striped state is unstable to transverse fluctuations that spoil the striped pattern if the randomizing event is delayed much beyond the time of Fig. 3b. This is shown in Fig. 3c, d, where we display snapshots of the continued evolution of the two-stripe state after successive, equal increments of time. The root cause of the instability in this particular system is that conservation of momentum guarantees that mean particle velocities parallel to the stripes will persist after collisions have run their course. Consequently, clusters form along newly spawned stripes; these clusters are evident both in Fig. 3d and in the upper left inset to Fig. 2b.

So when particles are randomized at periods close to $\tau = D/V$, transverse fluctuations do not have time to grow and the striped state is stable. But if randomizing events come much further apart than τ , this state loses stability to a clustered state. One such state is the square pattern in Fig. 2e–h, which shows a time series after noise has been added instantaneously to velocities of 4,000 particles clustered near the corners of a square. In this case, particles again explode away from their initial locations (Fig. 3e), collide inelastically along the symmetry lines separating the initial clusters (Fig. 3f), and then form new clusters at locations centred between the former clusters (Fig. 3g, h). If the plane is tiled with clusters located at the gridpoints of a square lattice, this state can reinforce itself, and so the very effect that destabilizes the striped pattern can stabilize the square pattern.

In this model system, the form of the particular instability that leads from one patterned state to another is heuristically transparent. In other problems, the instability will differ and may be more difficult to determine. In all pattern-formation problems, whether associated with noise or not, reflection, translation and rotation symmetries can be found to underlie the overt pattern. Thus the striped state is associated with symmetries of reflection along lines (dashed line in Fig. 3a) separating the initial stripes and of translation along the stripe direction. The square state is associated with reflection about two perpendicular lines (dashed lines in Fig. 3e). The square state appears exactly when the translational symmetry of the striped state is broken. Without these symmetries (as in problems where no interlaced dual pattern exists because, for example, periodic driving is absent^{41,53}) clustering can occur as a result of dissipation, but periodic pattern formation will not.

When do disordered states occur?

Spontaneously organized patterns can be formed by combining little more than noise (or diffusion) with dissipation (or nonlinear reactions) in a symmetric system. Why is it, then, that in some problems — such as the rubbing of a blackboard with flannel — a disordered, homogeneous state results, whereas in similar circumstances — the rubbing of a balloon with the same flannel — ordered, heterogeneous states appear⁵⁴?

To address this question, we paradoxically again turn to entropy. We have mentioned that the second law guarantees that the continual infusion of energy needed to maintain nonequilibrium states such as those shown in Figs 1 and 2a will always be sufficient to increase the entropy of the Universe. This coarse calculation is reassuring, but more can be learned by evaluating the entropy on a microcanonical basis — that is, on probabilistic grounds, one can conclude that the physically realized states that nature chooses tend to correspond to the ones that maximize the number of possible particle rearrangements. By counting all possible states, one finds that so-called ‘entropic ordering’ can favour either heterogeneous or homogeneous states with a transition at calculable values of order parameters, such as constituent concentrations.

As an example, suppose that we calculate the number of possible rearrangements of large and small particles constrained so that particle centres must lie on a one-dimensional grid⁵⁵. In Fig. 4 we sketch a much-simplified case where statistics of the entire microcanonical

Box 2

Liquid-crystal patterns

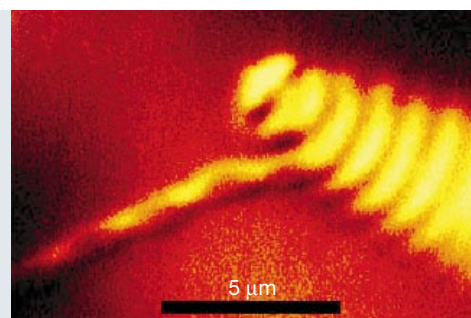
Colloidal ordering has been known for many years to be governed by entropy⁵⁴. This is important in several technological problems, including the generation of liquid-crystal displays and the fabrication of polymer composites. Intriguing examples in biological systems have also been found. As an example, the figure below displays an intermediate stage in the kinetics of the isotropic–smectic transition of a polymer/virus mixture. Initially, droplets of the metastable nematic phase are formed out of the isotropic background. Next, a single layer of the smectic phase is nucleated on the surface of the nematic drop. Growth proceeds by coalescence of similar drops until large droplets wrapped in smectic layers are formed. When a droplet is fully wound with smectic layers, a smectic strand peels away from the droplet, as shown in the figure. The strand has a uniform diameter of one-rod length. The rods are oriented perpendicular to the axis of the strand and rotate in this plane as seen by the modulations in intensity. The strands grow hundreds of microns long, unwind, and settle to the bottom of the container. Over a period of several months they transform into large, single-layer smectic sheets. These then stack on top of one other and eventually a single-phase, multilayer smectic monolith is formed.

Box 2 Figure

Entropic ordering of viral particles.

A tactoid of spontaneously aggregated rod-like viral particles. This micrograph was taken using a $\times 60$ water immersion lens

and differential-interference contrast optics and was subsequently false coloured. (Photo: Z. Dogic and S. Fraden, Brandeis University.)



ensemble can be evaluated using simple arithmetic. If we choose a representative example in which the large particles have radii just over 1.5 grid spacings, then these particles can exclude other particles from occupying adjacent gridpoints. This is indicated in Fig. 4, where the large grey particles are surrounded by excluded gridpoints indicated by red x's. The small blue particles can be arranged in $R = g!/s!(g-s)!$ distinct configurations (that is, excluding rearrangements of identical particles), where g is the number available gridpoints and s is the number of small particles. Notice that the volume excluded around the large particles increases either when the large particles are separated or when the large particles stray away from the edges of the grid. Correspondingly, in either of these situations the number of possible rearrangements of the small particles plunges from $R = 21$ to $R = 1$. This 21-fold change in the number of possible rearrangements implies that, just based on counting, we can conclude that it is highly probable that a randomly placed state will have large particles near one another, and that it is even more probable that the large particles will be near the edges^{56,57}.

If we calculate all possible rearrangements of both large and small particles, which is possible in this tiny illustrative problem, we find 154 distinct ‘separated’ states where all of one or the other species is contiguous, and only 116 ‘mixed’ states having at least one different particle intervening between like particles. Of the separated states, all but 24 have one particle adjacent to an edge. Thus there is a higher likelihood that we will find randomly distributed particles in a separated state than in a mixed state, and by far the highest likelihood

is that the large particles will cluster beside an edge. Notice that once again, irrespective of the specifics of any given problem, *any* system that is subject to random (that is, noisy) rearrangements will tend preferentially towards these separated states.

States in which particles separate into heterogeneous arrangements can be significantly favoured in randomized systems. In larger, more realistic problems, the probability contrast becomes much more extreme and separated states can be enormously more probable than any mixed state. This enhanced probability of obtaining separation due to 'volume-exclusion' effects either in the bulk^{58,59} or near boundaries^{56,57} persists in continuous^{56–64} as well as in simplified lattice-based examples, and has been reported widely in a variety of practical colloidal systems (see Box 2 for a topical example). But notice that this mechanism depends crucially on the concentrations of the two species of particles. By changing concentrations, one can tune the relative likelihood of obtaining a separated or a mixed state. This fact has been exploited in practical problems. For example, simulations^{58,60–64} and experiments¹¹ have confirmed that polymers of various shapes and sizes separate into distinct phases precisely when statistical calculations indicate that separation is favoured entropically over mixing. This result is of technological importance, for it implies that patterned displays, nanostructures and polymer blends can be engineered scientifically from first principles^{11,56}.

Using the volume-exclusion model, we can serendipitously directly compare microscale, entropically governed phase separation, observed in colloidal and polymeric systems, and macroscale noise-dominated segregation^{59,65}, long known to be ubiquitous (and troublesome) in granular systems⁶⁶. It is notable in this regard that there is little difference between research methods, results, or directions on phase separations of polymers and colloids and on segregation of grains — similar entropic considerations, scaling laws and even hard-sphere simulations have been used in the two fields.

As a test case, therefore, we make alternative use of our previous observation that thin vibrated beds of grains exhibit effective randomization of their velocities by examining how segregation between different size grains changes as relative concentrations of the grains are varied. Results of such an experiment are presented in Fig. 5. Here we have vibrated a 5-mm-deep blend of large (gold) and small (maroon) glass beads under identical conditions, but at three different sets of concentrations. At high concentration of small beads (Fig. 5a), the small and large beads separate cleanly, each species favouring positions near walls. As the relative concentration of small beads is reduced, the entropic advantage of mixing grows with respect to that of separation, and only a small region (bottom right of Fig. 5b) of segregated small beads is seen. Finally, the entropy of the mixed state exceeds that of the homogeneous state, which prevails when the weight fraction of small beads in the bed drops below about 11% (Fig. 5c). We cannot rule out the possibility that the small beads are in fact segregated, perhaps forming an island hidden by the larger grains, but if such islands exist, they are not visible and they do not change the bed dynamics in any perceptible way. Coincidentally, the fraction of smaller beads at which the transition between mixing and separation appears is close to the transition concentration predicted using hard-sphere simulations applied to polymers⁶³.

This system is far from the equilibrium case considered using traditional statistical mechanics. Moreover, the microscale physics of polymer blends, viral crystal aggregates and other systems, kept under controlled temperature and chemical balance, are quite different from those of large, energetically vibrated glass beads. Nevertheless the granular system is highly randomized and *because of this fact*, the probabilistic considerations that favour macroscale patterned structures over homogeneous mixed states apply irrespective of these differences in detail.

Outlook

It has been said that any sufficiently advanced technology is indistinguishable from magic. The inseparable facts that continue to attract

researchers to the study of patterns generated from noise are that the mechanism for their formation is extremely basic, yet even once the mechanism has been thoroughly dissected and understood, its outcome still seems like magic. The paradox that the spontaneous formation of these patterns seems simultaneously to be completely implausible and hugely probable is difficult to reconcile with a rational and systematic analysis of the natural world. This situation has given rise to several unresolved, and possibly unresolvable, questions.

For example, if random processes generate similar patterns to those generated by biological systems, is there any way to interpret patterns to determine which of these causes have produced putative fossil remains? In one sense this is a religious question: are the patterns that constitute life itself possible through only random influences, or was a deterministic, intelligent intervention involved? At this level, the question may be unresolvable, but on another front, several researchers^{49,67–69} have developed concrete relations between pattern selection and underlying symmetries. What are the practical lessons that this research can teach us about patterns and the mechanisms that formed them? More philosophically, in statistical physics, profound debates have developed historically over the direction of time — which is often characterized based on the tendency of ordered states become disordered over time. These debates have so far been partially resolved by establishing that the tendency toward disorder is due to its vastly greater probability of occurrence²⁶. How shall we interpret these debates when ordered states are actually more probable?

Finally, as research becomes directed increasingly towards pressing technological issues in problems such as nanofabrication, smart materials, semiconductor design and thin-film coating, to name a few, the engineering of spontaneously assembling patterns has begun to come into its own, and the question of potential limits to the design of entropically controlled patterns arises. For example, just using the model described in Fig. 2 that was designed deliberately to generate patterns using noise and dissipation alone, we have seen that squares, stripes, hexagons and triangles can be generated. The first three of these have been observed experimentally, but so far the triangles have not. Yet they respect perfectly well defined symmetries — does this mean that there are additional, currently unknown constraints preventing the physical manifestation of certain mathematically allowed patterns? To the triangles example, one can add an enormous variety of other patterns obtained from symmetry analysis, including quasipatterns^{70,71}, superlattices^{72,73} and three-dimensional patterns^{40,74,75}. Understanding the natural limits on pattern selection remains an on-going challenge. □

- Shinbrot, T. A novel model of contact charging. *J. Electrostat.* **17**, 113–123 (1985).
- Ball, P. *The Self-made Tapestry: Pattern Formation in Nature* (Oxford Univ. Press, Oxford, 1999).
- Graham, M. D. et al. Effects of boundaries on pattern formation: catalytic oxidation of CO on platinum. *Science* **264**, 80–82 (1994).
- Vogel, G. Goodness gracious, great walls afar. *Science* **274**, 343–343 (1996).
- Toomre, A. & Toomre, J. Galactic bridges and tails. *Astrophys. J.* **178**, 623–666 (1972).
- Sellwood, J. A. Spiral structure as a recurrent instability. *Astrophys. Space Sci.* **272**, 31–43 (2000).
- Kondo, S. & Asai, R. A reaction-diffusion wave on the skin of the marine angelfish *Pomacanthus*. *Nature* **376**, 765–768 (1995).
- Haim, D. et al. Breathing spots in a reaction-diffusion system. *Phys. Rev. Lett.* **77**, 190–193 (1996).
- Pearson, J. E. Complex patterns in a simple system. *Science* **261**, 189–192 (1993).
- Kadar, S., Wang, J. & Showalter, K. Noise supported traveling waves in sub-excitable media. *Nature* **391**, 770–772 (1998).
- Adams, M., Dogic, Z., Keller, S. L. & Fraden, S. Entropically driven microphase transitions in mixtures of colloidal rods and spheres. *Nature* **393**, 349–352 (1998).
- Grotzinger, J. P. & Rothman, D. H. An abiotic model for stromatolite morphogenesis. *Nature* **383**, 423–425 (1996).
- Csahók, I., Z., Misbah, I., Rioual, F. & Valance, A. Dynamics of aeolian sand ripples. *Eur. Phys. J. E* **3**, 71–86 (2000).
- Makse, H. A. Grain segregation mechanisms in aeolian sand ripples. *Eur. Phys. J. E* **1**, 127–135 (2000).
- Lekkerkerker, H. N. W. & Stroobants, A. Ordering entropy. *Nature* **393**, 305–306 (1998).
- Reynolds, O. *On the Action of Rain to Calm the Sea. Papers on Mechanical and Physical Subjects* 86–88 (Cambridge Univ. Press, London, 1900).
- Tsimplis, M. N. The effect of rain in calming the sea. *J. Phys. Oceanogr.* **22**, 404–412 (1993).
- Herzel, H. Stabilization of chaotic orbits by random noise. *A. Angew. Math. Mech.* **68**, 11–12 (1988).
- Walker, J. *The Flying Circus of Physics* 14–8, 99–100 (Wiley, New York, 1975).
- Weisenfeld, K. & Moss, F. Stochastic resonance and the benefits of noise: from ice ages to crayfish and SQUIDS. *Nature* **373**, 33–36 (1995).
- Barnsley, M. *Fractals Everywhere* (Academic, Boston, 1988).

22. Jung, P. & Mayer-Kress, G. Spatiotemporal stochastic resonance in excitable media. *Phys. Rev. Lett.* **74**, 2130–2133 (1995).
23. Shinbrot, T. & Scarbrough, K. Using variability to regulate long term biological rhythms. *J. Theor. Biol.* **196**, 455–471 (1999).
24. Strogatz, S. H. *Nonlinear Dynamics and Chaos* 254–255 (Addison Wesley, Reading, MA, 1994).
25. Caves, C. M., Unruh, W. G. & Zurek, W. H. Comment on 'Quantitative limits on the ability of a Maxwell demon to extract work from heat'. *Phys. Rev. Lett.* **65**, 1387 (1990).
26. Brush, S. G. *Statistical Physics and the Atomic Theory of Matter* 86–90 (Princeton Univ. Press, Princeton, 1983).
27. Schlichting, H. J. & Nordmeier, V. Strukturen im Sand—Kollektives Verhalten und Selbstorganisation bei Granulaten Math. *Naturwissenschaften* **49**, 323–332 (1996).
28. Eggers, J. Sand as Maxwell's demon. *Phys. Rev. Lett.* **83**, 5322–5325 (1999).
29. Sano, O. Random motion of a marker particle on square cells formed on vertically vibrated granular layer. *J. Phys. Soc. Jpn* **68**, 1769–1777 (1999).
30. Shinbrot, T. Competition between randomizing impacts and inelastic collisions in granular pattern formation. *Nature* **389**, 574–576 (1997).
31. Bizon, C., Shattuck, M. D., Swift, J. B. & Swinney, H. L. Transport coefficients for granular media from molecular dynamics simulations. *Phys. Rev. E* **60**, 4340–4351 (1999).
32. Winfree, A. Heart muscle as a reaction-diffusion medium: the roles of electric potential, diffusion, activation front curvature, and anisotropy. *Int. J. Bif. Chaos* **7**, 487–526 (1997).
33. Winfree, A. Electrical turbulence in 3-dimensional heart muscle. *Science* **266**, 1003–1006 (1994).
34. Klausmeier, C. A. Regular and irregular patterns in semiarid vegetation. *Science* **284**, 1826–1828 (1999).
35. Lee, K. J., McCormick, W. D., Ouyang, Q. & Swinney, H. L. Pattern formation by interacting chemical fronts. *Science* **261**, 192–194 (1993).
36. Gorman, M., el-Hamdi, M. & Pearson, B. Ratcheting motion of concentric rings in cellular flames. *Phys. Rev. Lett.* **76**, 228–231 (1996).
37. Aldushin, A. P. & Matkowsky, B. J. *Diffusion Driven Combustion Waves in Porous Media* (Gordon & Breach, Newark, NJ, 1999).
38. Sellwood, J. A. & Carlberg, R. G. Spiral instabilities provoked by accretion and star formation. *Astrophys. J.* **282**, 61–74 (1984).
39. Wu, K. K. S., Lahav, O. & Rees, M. J. The large-scale smoothness of the universe. *Nature* **397**, 225–230 (1999).
40. Einasto, J. *et al.* A 120-mpc periodicity in the three-dimensional distribution of galaxy superclusters. *Nature* **385**, 139–141 (1997).
41. Goldhirsch, I. & Zanetti, G. Clustering instability in dissipative gases. *Phys. Rev. Lett.* **70**, 1619–1622 (1993).
42. Bizon, C., Shattuck, M. D., Swift, J. B., McCormick, W. D. & Swinney, H. L. Patterns in 3D vertically oscillated granular layers: simulation and experiment. *Phys. Rev. Lett.* **80**, 57–60 (1998).
43. Shinbrot, T., Lomelo L. & Muzzio, F. J. Harmonic patterns in fine granular vibrated beds. *Gran. Matt.* **2**, 65–69 (2000).
44. Umbanhowar, P. B., Melo, F. & Swinney, H. L. Localized excitations in a vertically vibrated granular layer. *Nature* **382**, 793–796 (1996).
45. Eggers, J. & Riecke, H. A continuum description of vibrated sand. *Phys. Rev. E* **59**, 4476–4483 (1999).
46. Tsimring, L. S. & Aranson, I. S. Localized and cellular patterns in a vibrated granular layer. *Phys. Rev. Lett.* **79**, 213–216 (1997).
47. Venkataramani, S. C. & Ott, E. Spatio-temporal bifurcation phenomena with temporal period doubling: patterns in vibrated sand. *Phys. Rev. Lett.* **80**, 3495–3498 (1998).
48. Metcalf, T. H., Knight, J. B. & Jaeger, H. M. Standing wave patterns in shallow beds of vibrated granular material. *Physica A* **236**, 202–210 (1997).
49. Crawford, J. D., Gollub, J. P. & Lane, D. Hidden symmetries of parametrically forced waves. *Nonlinearity* **6**, 119–164 (1993).
50. Sinai, Ya. G. Dynamical systems with elastic reflections: ergodic properties of dispersing billiards. *Russ. Math. Surv.* **25**, 137–147 (1970).
51. Ding, M.-Z., Grebogi, C., Ott, E. & Yorke, J. A. Transition to chaotic scattering. *Phys. Rev. A* **42**, 7025–7040 (1990).
52. Luding, S., Clément, E., Blumen, A., Rajchenbach, J. & Duran, J. Studies of columns of beads under external vibrations. *Phys. Rev. E* **49**, 1634–1646 (1994).
53. Aranson, I. S. *et al.* Electrostatically driven granular media: phase transitions and coarsening. *Phys. Rev. Lett.* **84**, 3306–3309 (2000).
54. Grier, D. G. A surprisingly attractive couple. *Nature* **393**, 621–622 (1998).
55. Gujrati, P. D. Entropy-driven phase separation and configurational correlations on a lattice: some rigorous results. *Phys. Rev. E* (in the press).
56. Dinsmore, A. D., Yodh, A. G. & Pine, D. J. Entropic control of particle motion using passive surface microstructures. *Nature* **383**, 239–242 (1996).
57. Crocker, J. C. & Grier, D. G. When like charges attract: the effects of geometrical confinement on long-range colloidal interactions. *Phys. Rev. Lett.* **77**, 1897–1900 (1996).
58. Crocker, J. C., Matteo, J. A., Dinsmore, A. D. & Yodh, A. G. Entropic attraction and repulsion in binary colloids probed with a line optical tweezer. *Phys. Rev. Lett.* **82**, 4352–4355 (1999).
59. Duran, J. & Jullien, R. Attractive forces in a granular cocktail. *Phys. Rev. Lett.* **80**, 3547–3550 (1998).
60. van Roij, R., Mulder, B. & Dijkstra, M. Phase behavior of binary mixtures of thick and thin hard rods. *Physica A* **261**, 374–390 (1998).
61. Dijkstra, M. & Frenkel, D. Evidence for entropy-driven demixing in hard-core fluids. *Phys. Rev. Lett.* **72**, 298–300 (1994).
62. Dijkstra, M. & van Roij, R. Entropy-driven demixing in binary hard-core mixtures: from hard spherocylinders towards hard spheres. *Phys. Rev. E* **56**, 5594–5602 (1997).
63. Dijkstra, M., van Roij, R. & Evans, R. Phase diagram of highly asymmetric binary hard-sphere mixtures. *Phys. Rev. E* **59**, 5744–5771 (1999).
64. König, A. & Ashcroft, N. W. Structure and effective interactions three-component hard sphere liquids. *Phys. Rev. E* (in the press).
65. Hong, D., Quinn, P. V. & Luding, S. Reverse Brazilian nut problem: competition between percolation and condensation. *Phys. Rev. Lett.* (in the press); preprint cond-mat/0010459 also available at <http://xxx.lanl.gov>.
66. Shinbrot, T. & Muzzio, F. J. Nonequilibrium patterns in granular mixing and segregation. *Phys. Today* 25–30 (March 2000).
67. Stewart, I. & Golubitsky, M. *Fearful Symmetry: Is God a Geometer?* (Penguin, London, 1992).
68. Silber, M. & Skeldon, A. C. Parametrically excited surface waves: two-frequency forcing, normal form symmetries, and pattern selection. *Phys. Rev. E* **59**, 5446–5456 (1999).
69. Bowden, N., Terfort, A., Carbeck, J. & Whitesides, G. M. Self-assembly of mesoscale objects into ordered two-dimensional arrays. *Science* **276**, 233–236 (1997).
70. Silber, M., Topaz, C. M. & Skeldon, A. C. Two-frequency forced Faraday waves: weakly damped modes and pattern selection. *Physica D* **143**, 205–225 (2000).
71. Edwards, W. S. & Fauve, S. Patterns and quasi-patterns in the Faraday experiment. *J. Fluid Mech.* **278**, 123–148 (1994).
72. Kudrolli, A., Pier, B. & Gollub, J. P. Superlattice patterns in surface waves. *Physica D* **22**, 99–111 (1998).
73. Silber, M. & Proctor, M. R. E. Nonlinear competition between small and large hexagonal patterns. *Phys. Rev. Lett.* **81**, 2450–2453 (1998).
74. van Blaaderen, A., Ruel, R. & Wiltzius, P. Template-directed colloidal crystallization. *Nature* **385**, 321–323 (1997).
75. Prause, B., Glazier, J. A., Gravina, S. & Montemagno, C. Magnetic resonance imaging of a three dimensional foam. *J. Phys.* **7**, L511–L516 (1995).
76. Bagnold, R. A. *The Physics of Blown Sand and Desert Dunes* (Methuen, London, 1941).
77. Forrest, S. B. & Haff, P. K. Mechanics of wind ripple stratigraphy. *Science* **255**, 1240–1243 (1992).
78. Heaney, P. J. & Davis, A. M. Observation and origin of self-organized textures in agates. *Science* **269**, 1562–1565 (1995).
79. Hosoi, A. E. & Dupont, T. F. Layer formation in monodisperse suspensions and colloids. *J. Fluid Mech.* **328**, 297–311 (1996).
80. Mueth, D. M., Crocker, J. C., Esipov, S. E. & Grier, D. G. Origin of stratification in creaming emulsions. *Phys. Rev. Lett.* **77**, 578–581 (1996).
81. Mullin, T. Coarsening of self-organized clusters in binary mixtures of particles. *Phys. Rev. Lett.* **84**, 4741–4744 (2000).
82. Oliver, S., Kuperman, A., Coombs, N., Lough, A. & Ozin, G. A. Lamellar aluminophosphates with surface patterns that mimic diatom and radiolarian microstructures. *Nature* **378**, 47–50 (1995).
83. Dogic, Z. & Fraden, S. Smectic phase in a colloidal suspension of semiflexible virus particles. *Phys. Rev. Lett.* **78**, 2417–2420 (1997).
84. Kestenbaum, D. Gentle force of entropy bridges disciplines. *Science* **279**, 1849–1849 (1998).

Acknowledgements

We acknowledge helpful remarks and vital contributions from M. Dijkstra, Z. Dojic, S. Fraden, P. Garik, R. van Roij, H. Swinney and P. Umbanhowar.

Supercooled liquids and the glass transition

Pablo G. Debenedetti* & Frank H. Stillinger†‡

*Department of Chemical Engineering and ‡Princeton Materials Institute, Princeton University, Princeton, New Jersey 08544, USA (e-mail: pdebene@princeton.edu)

†Bell Laboratories, Lucent Technologies, Murray Hill, New Jersey 07974, USA

Glasses are disordered materials that lack the periodicity of crystals but behave mechanically like solids. The most common way of making a glass is by cooling a viscous liquid fast enough to avoid crystallization. Although this route to the vitreous state — supercooling — has been known for millennia, the molecular processes by which liquids acquire amorphous rigidity upon cooling are not fully understood. Here we discuss current theoretical knowledge of the manner in which intermolecular forces give rise to complex behaviour in supercooled liquids and glasses. An intriguing aspect of this behaviour is the apparent connection between dynamics and thermodynamics. The multidimensional potential energy surface as a function of particle coordinates (the energy landscape) offers a convenient viewpoint for the analysis and interpretation of supercooling and glass-formation phenomena. That much of this analysis is at present largely qualitative reflects the fact that precise computations of how viscous liquids sample their landscape have become possible only recently.

The glassy state is ubiquitous in nature and technology¹. It is crucial in the processing of foods², the commercial stabilization of labile biochemicals³, and the preservation of insect life under extremes of cold or dehydration³.

Window glass, composed mostly of sand, lime and soda, is the best-known example of an engineered amorphous solid⁴. Optical fibres are made of very pure amorphous silica, occasionally carefully doped. Most engineering plastics are amorphous solids, as are some metallic glasses and alloys of interest because of their soft magnetism and corrosion resistance⁵. The silicon used in many photovoltaic cells is amorphous, and it is possible that most water in the Universe may be glassy⁶. Most of these examples entail supercooling of a liquid to take advantage of viscous retardation of nucleation and crystallization. Understanding quantitatively the extraordinary viscous slow-down that accompanies supercooling and glass formation is a major scientific challenge⁷.

We begin by reviewing the phenomenology of vitrification and supercooling. A useful approach for relating this complex phenomenology to molecular-scale events is to focus attention on the liquid's energy landscape, that is, the multidimensional surface generated by the system's potential energy as a function of molecular coordinates. Accordingly, basic landscape concepts and a discussion of the important theoretical and computational progress currently being made in this area are presented next. This is followed by a discussion of alternative viewpoints, in which narrowly avoided singularities are assumed to occur well above the glass-transition temperature. We then close with a summary of the important open questions.

It is impossible to do justice to the entire field of supercooled liquids and amorphous solids in an article of this length. We have therefore limited the scope to the dynamics and thermodynamics of viscous liquids above and close to the glass-transition temperature T_g — in other words, to the glass transition viewed 'from the liquid'. The view 'from the solid', including such topics as relaxation both relatively near

(for example, during annealing or ageing) and far below T_g , is not discussed. The reader is referred to an excellent recent review⁸ for a thorough coverage of these and other topics.

Phenomenology of supercooling and glass formation

Figure 1 illustrates the temperature dependence of a liquid's volume (or enthalpy) at constant pressure^{4,9}. Upon cooling below the freezing point T_m , molecular motion slows down. If the liquid is cooled sufficiently fast, crystallization can be avoided^{10,11}. Eventually molecules will rearrange so slowly that they cannot adequately sample configurations in the available time allowed by the cooling rate. The liquid's structure therefore appears 'frozen' on the laboratory timescale (for example, minutes). This falling out of equilibrium occurs across a narrow transformation range where the characteristic molecular relaxation time becomes of the order of 100 seconds, and the rate of change of volume or enthalpy with respect to temperature decreases abruptly (but continuously) to a value comparable to that of a crystalline solid. The resulting material is a glass. The intersection of the liquid and vitreous portions of the volume versus temperature curve provides one definition of T_g , which usually occurs around $2T_m/3$. The behaviour depicted in Fig. 1 is not a true phase transition, as it does not involve discontinuous changes in any physical property.

The slower a liquid is cooled, the longer the time available for configurational sampling at each temperature, and hence the colder it can become before falling out of liquid-state equilibrium. Consequently, T_g increases with cooling rate^{12,13}. The properties of a glass, therefore, depend on the process by which it is formed. In practice, the dependence of T_g on the cooling rate is weak (T_g changes by 3–5 °C when the cooling rate changes by an order of magnitude¹⁴), and the transformation range is narrow, so that T_g is an important material characteristic.

Slowing down

Another definition of T_g is the temperature at which the shear viscosity reaches 10^{13} poise. Close to T_g the viscosity η

is extraordinarily sensitive to temperature. For silica this dependence is reasonably well described by the Arrhenius functionality, $\eta = A \exp(E/k_B T)$, where A and E are temperature-independent and k_B is Boltzmann's constant. Other liquids exhibit an even more pronounced viscous slow-down close to the glass transition, which is reasonably well represented, over 2–4 orders of magnitude in viscosity⁸, by the Vogel–Tammann–Fulcher (VTF) equation^{15–17}

$$\eta = A \exp[B/(T - T_0)] \quad (1)$$

where A and B are temperature-independent constants. Understanding the origin of this extraordinary slow-down of relaxation processes is one of the main challenges in the physics of glasses.

Figure 2 shows a T_g -scaled Arrhenius representation of liquid viscosities^{18–21}. Angell has proposed a useful classification of liquids

along a 'strong' to 'fragile' scale. The viscosity and relaxation times (for example, dielectric relaxation) of the former behave in nearly Arrhenius fashion, whereas fragile liquids show marked deviations from Arrhenius behaviour. Silica (SiO_2) is the prototypical strong liquid, whereas *o*-terphenyl (OTP) is the canonical fragile glass-former. Strong liquids, such as the network oxides SiO_2 and germanium dioxide (GeO_2), have tetrahedrally coordinated structures, whereas the molecules of fragile liquids exert largely non-directional, dispersive forces on each other. Alternative scaling descriptions that attempt to extract universal aspects of viscous slow-down have been proposed^{22–24}. Their relative merits are still being assessed^{8,25}.

Viscous liquids close to T_g exhibit non-exponential relaxation. The temporal behaviour of the response function $F(t)$ (for example, the polarization in response to an applied electric field, the strain (deformation) resulting from an applied stress, or the stress in

Box 1

Entropy crises

Boltzmann's entropy formula establishes the connection between the microscopic world of atoms and molecules and the bulk properties of matter:

$$S(N, V, E) = k_B \ln \Omega$$

In this equation, S is the entropy, k_B is Boltzmann's constant and Ω is the number of quantum states accessible to N particles with fixed energy E in a volume V . Because Ω cannot be less than one, the entropy cannot be negative. When a crystal is cooled sufficiently slowly, it approaches a unique state of lowest energy, and hence its entropy approaches zero as $T \rightarrow 0$. If the entropy of a supercooled liquid were to become smaller than that of the stable crystal at the Kauzmann temperature, its entropy would eventually become negative upon further cooling. This impossible scenario constitutes an entropy crisis^{46–48}.

The Kauzmann temperature T_K is given by⁹

$$\Delta S_m = \int_{T_K}^{T_m} \frac{\Delta C_p}{T} dT$$

where ΔS_m is the melting entropy (the difference between liquid and crystal entropies at the melting temperature), T_m is the melting temperature at the given pressure, and ΔC_p is the temperature-dependent difference between the heat capacity of the liquid and the crystal at the given pressure. The rate of change of entropy with temperature at constant pressure is given by

$$\left(\frac{\partial S}{\partial T}\right)_P = \frac{C_p}{T}$$

The entropy crisis arises because the heat capacity of a liquid is greater than that of the stable crystal. The entropy of fusion is therefore consumed upon supercooling, and vanishes at T_K . The entropy crisis entails no conflict with the second law of thermodynamics, as the difference in chemical potential $\Delta\mu$ between the supercooled liquid and the stable crystal at T_K is a positive quantity. Because the chemical potential is the Gibbs free energy per unit mass, this means that the system can reduce its Gibbs free energy by freezing, in accord with experience. The chemical potential difference at T_K is given by⁹

$$\Delta\mu(T_K) = \int_{T_K}^{T_m} \Delta C_p \left(\frac{T_m}{T} - 1\right) dT$$

One way of avoiding the entropy crisis is for the liquid to form an ideal glass of unique configuration at T_K . This is the thermodynamic view of the glass transition, according to which the observable glass transition is a manifestation, masked by kinetics, of an underlying second-order phase transition⁵⁰ occurring at T_K .

Because the glass transition intervenes before the entropy crisis occurs ($T_g > T_K$), estimates of the Kauzmann temperature involve an extrapolation of liquid properties below T_g . The validity of such extrapolations, and hence of the very possibility of an entropy crisis, has been questioned by Stillinger¹⁰³, owing to the apparent necessity for configurational excitations out of the ideal glass state to require an unbounded energy. Furthermore, recent computer simulations of polydisperse hard disks found no evidence of a thermodynamic basis underlying the glass transition¹⁰⁴. Although the notion of an ideal glass remains controversial¹⁰³, this does not undermine the usefulness of T_K as an empirical classification parameter for glass-formers.

Experimentally, there are substances with known Kauzmann points. ⁴He is a liquid at 0 K and 1 bar (liquid He-II). Its equilibrium freezing pressure at 0 K is 26 bar. At this point, the entropies of the liquid and the crystal are equal, and this is therefore a Kauzmann point, although not an entropy crisis as both phases have zero entropy. The melting curves of ³He and ⁴He exhibit pressure minima: these occur at about 0.32 K and 0.8 K, respectively¹⁰⁵. These are also equal-entropy (Kauzmann) points. Experiments indicate that poly(4-methylpentene-1) exhibits a pressure maximum along its melting curve¹⁰⁶. Although the appearance of an additional phase complicates the interpretation of this observation, the implication would be that the pressure maximum is a Kauzmann point, and the continuation of the melting curve to lower temperatures and pressures beyond this point corresponds to endothermic freezing of a stable liquid into a crystal possessing higher entropy¹⁰⁷. How this liquid would avoid conflict with the third law is not understood, but may hinge on the vibrational anharmonicity of the two phases with changing temperature.

response to an imposed deformation) can often be described by the stretched exponential, or Kohlrausch–Williams–Watts (KWW) function^{26,27}

$$F(t) = \exp[-(t/\tau)^\beta] \quad (\beta < 1) \quad (2)$$

where $F(t) = [\sigma(t) - \sigma(\infty)]/[\sigma(0) - \sigma(\infty)]$ and σ is the measured quantity (for example, the instantaneous stress following a step change in deformation). τ in equation (2) is a characteristic relaxation time, whose temperature dependence is often non-Arrhenius (exhibiting fragile behaviour). The slowing down of long-time relaxation embodied in equation (2) contrasts with the behaviour of liquids above the melting point, which is characterized by simple exponential relaxation. Experimental and computational evidence indicates that this slow-down is related to the growth of distinct relaxing domains^{28–39} (spatial heterogeneity). Whether each of these spatially heterogeneous domains relaxes exponentially or not is a matter of considerable current interest^{38,39}.

Decouplings

In supercooled liquids below approximately $1.2T_g$ there occurs a decoupling between translational diffusion and viscosity, and between rotational and translational diffusion^{30,39,40}. At higher temperatures, both the translational and the rotational diffusion coefficients are inversely proportional to the viscosity, in agreement with the Stokes–Einstein and Debye equations, respectively. Below approximately $1.2T_g$, the inverse relationship between translational motion and viscosity breaks down, whereas that between rotational motion and viscosity does not. Near T_g , it is found that molecules translate faster than expected based on their viscosity, by as much as two orders of magnitude. This therefore means that, as the temperature is lowered, molecules on average translate progressively more for every rotation they execute. Yet another decoupling occurs in the moderately supercooled range. At sufficiently high temperature the liquid shows a single peak relaxation frequency (Fig. 3), indicative of one relaxation mechanism. In the moderately supercooled regime, however, the peak splits into slow (α) and fast (β) relaxations^{41–43}. The former exhibit non-Arrhenius behaviour and disappear at T_g ; the latter continue below T_g and display Arrhenius behaviour⁴⁴.

Thermodynamics

The entropy of a liquid at its melting temperature is higher than that of the corresponding crystal. Because the heat capacity of a liquid is higher than that of the crystal, this entropy difference decreases upon supercooling (Box 1). Figure 4 shows the temperature dependence of the entropy difference between several supercooled liquids and their stable crystals⁴⁵. For lactic acid this entropic surplus is consumed so

fast that a modest extrapolation of experimental data predicts its impending vanishing. In practice, the glass transition intervenes, and ΔS does not vanish. If the glass transition did not intervene, the liquid entropy would equal the crystal's entropy at a nonzero temperature T_K (the Kauzmann temperature.) Because the entropy of the crystal approaches zero as T tends to zero, the entropy of the liquid would eventually become negative upon cooling if this trend were to continue. Because entropy is an inherently non-negative quantity (Box 1), the state of affairs to which liquids such as lactic acid are tending when the glass transition intervenes is an entropy crisis^{46–48}. The extrapolation needed to provoke conflict with the third law is quite modest for many fragile liquids⁴⁹, and the imminent crisis is thwarted by a kinetic phenomenon, the glass transition. This suggests a connection between the kinetics and the thermodynamics of glasses⁴⁷. The thermodynamic viewpoint that emerges from this analysis⁵⁰ considers the laboratory glass transition as a kinetically controlled manifestation of an underlying thermodynamic transition to an ideal glass with a unique configuration.

A formula of Adam and Gibbs⁵¹ provides a suggestive connection between kinetics and thermodynamics:

$$t = A \exp(B/Ts_c) \quad (3)$$

In this equation, t is a relaxation time (or, equivalently, the viscosity) and A and B are constants. s_c , the configurational entropy, is related to the number of minima of the system's multidimensional potential energy surface (Box 2). According to the Adam–Gibbs picture, the origin of viscous slow-down close to T_g is the decrease in the number of configurations that the system is able to sample. At the Kauzmann temperature the liquid would have attained a unique, non-crystalline state of lowest energy, the ideal glass. Because there is no configurational entropy associated with confinement in such a state, the Adam–Gibbs theory predicts structural arrest to occur at T_K . In their derivation of equation (3), Adam and Gibbs invoked the concept of a cooperatively rearranging region (CRR)⁵¹. A weakness of their treatment is the fact that it provides no information on the size of such regions. The fact that the CRRs are indistinguishable from each other is also problematic, in light of the heterogeneity that is believed to underlie stretched exponential behaviour⁸.

Figure 1 Temperature dependence of a liquid's volume v or enthalpy h at constant pressure.

T_m is the melting temperature. A slow cooling rate produces a glass transition at T_{ga} ; a faster cooling rate leads to a glass transition at T_{gb} .

The thermal expansion coefficient $\alpha_p = (\partial \ln v / \partial T)_p$ and the isobaric heat capacity $c_p = (\partial h / \partial T)_p$ change abruptly but continuously at T_g .

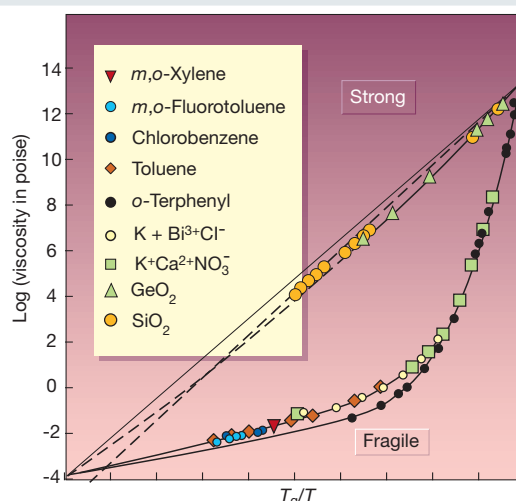
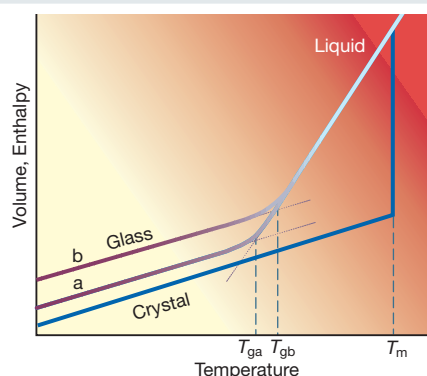


Figure 2 T_g -scaled Arrhenius representation of liquid viscosities showing Angell's strong–fragile pattern. Strong liquids exhibit approximate linearity (Arrhenius behaviour), indicative of a temperature-independent activation energy $E = \ln \eta / d(1/T) \approx \text{const}$. Fragile liquids exhibit super-Arrhenius behaviour, their effective activation energy increasing as temperature decreases. (Adapted from refs 9 and 11.)

Nevertheless, equation (3) describes the relaxation behaviour of deeply supercooled liquids remarkably well. If the difference in heat capacities between a supercooled liquid and its stable crystalline form is inversely proportional to temperature⁵², the Adam–Gibbs

relation yields the VTF equation, which is mathematically equivalent to the Williams–Landel–Ferry equation for the temperature dependence of viscosity in polymers⁵³. This transformation is predicated on the assumption that the vibrational entropies of the

Box 2

Statistics of landscapes

The complexity of many-body landscapes makes a statistical description inevitable. The quantity of interest is the number of minima of given depth, which is given by¹⁰⁸

$$\frac{d\Omega}{d\phi} = C \exp[N\sigma(\phi)]$$

Here, $d\Omega$ denotes the number of potential energy minima with depth per particle ($\phi = \Phi/N$) between ϕ and $\phi \pm d\phi/2$. C is an N -independent factor with units of inverse energy, and $\sigma(\phi)$, also an N -independent quantity, is a so-called basin enumeration function. Taking the logarithm of the above expression and comparing with Boltzmann's entropy formula (Box 1), we see that $\sigma(\phi)$ is the entropy per particle arising from the existence of multiple minima of depth ϕ , or, in other words, the configurational entropy.

At low temperatures, it is possible to separate the configurational contribution to thermophysical properties, which arises from the exploration of different basins, from the vibrational component, which arises from thermal motions confined to a given basin^{75,76}. The Helmholtz free energy A is then given by

$$\frac{A}{NkT} = \frac{\bar{\phi}}{kT} - \sigma(\bar{\phi}) + \frac{a^v}{k_B T}$$

where $\bar{\phi}$ is the depth of the basins preferentially sampled at the given temperature, and a^v is the vibrational free energy per particle. Thus, the free energy consists of an energetic component that reflects the depth of landscape basins sampled preferentially at the given temperature, an entropic component that accounts for the number of existing basins of a given depth, and a vibrational component. The statistical description of a landscape consists of the basin enumeration function $\sigma(\phi)$, from which the excitation profile $\phi(T)$ is obtained through the free-energy minimization condition

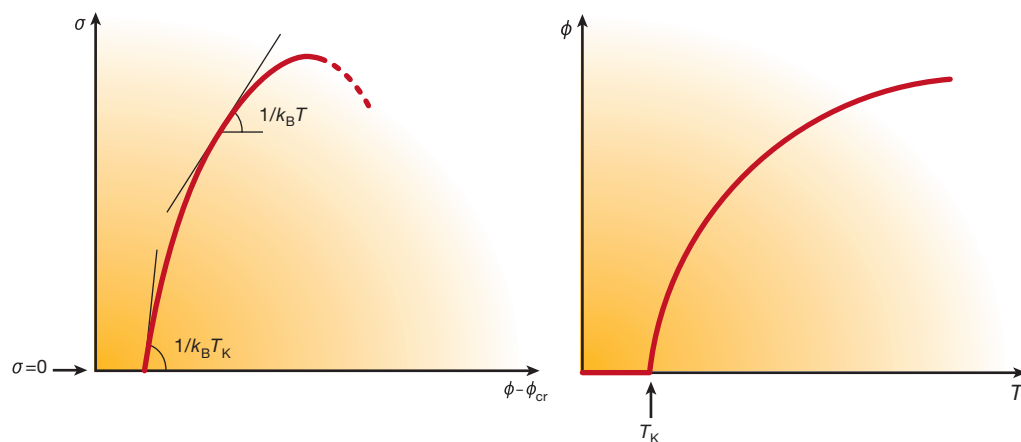
$$\frac{d\sigma}{d\phi} = \frac{1}{k_B T}$$

The above equation assumes that a^v depends on T , but not on ϕ — that is, all basins have the same mean curvature at their respective minima.

The shape of a given system's landscape is determined by the density (number of molecules per unit volume, N/V). Temperature governs the manner in which the landscape is sampled. A different basin enumeration function and excitation profile corresponds to each density. Temperature dictates the point along the enumeration curve and the excitation profile sampled by the system at fixed density (see figure below).

It is possible to construct the basin enumeration function and excitation profile of a system from experimental heat capacity data for the crystal and the supercooled liquid¹⁰⁹, and by computer simulation^{77,78}. In the latter case, the calculations involve determining the probability distribution of inherent structure energies sampled as a function of temperature. These calculations are at the limit of what is presently feasible with available computational power. The enumeration function is often well represented by a parabola, indicative of a gaussian distribution of basins^{4,77,97}. At present it is not understood how the enumeration function deforms with density for a given system (but see ref. 96 for a recent example of such a calculation), or how it depends on molecular architecture. Understanding such questions would provide a direct link between landscape statistics and physical properties. The success of the Adam–Gibbs equation indicates that this link applies also to transport properties such as diffusion and viscosity.

Box 2 Figure Schematic representation of the basin enumeration function (left) and the excitation profile (right). ϕ is the potential energy per particle in mechanically stable potential energy minima. ϕ_{cr} is the corresponding quantity in the stable crystal. The number of potential energy minima with depth between ϕ and $\phi \pm d\phi/2$ is proportional to $\exp[N\sigma(\phi)]$. In the thermodynamic limit (N of the order of Avogadro's number, 6.02×10^{23}), basins that possess larger (less negative) potential energies (shallow basins) are overwhelmingly more numerous than deeper basins possessing very negative ϕ -values. The slope of the enumeration function is inversely proportional to the temperature. The excitation profile gives the depth of the inherent structures sampled preferentially at a given temperature. At the Kauzmann temperature T_K the system attains the state of a configurationally unique ideal glass ($\sigma=0$), corresponding to the deepest amorphous basin (see Figs 5 and 8) and its inherent structure energy does not therefore change upon further cooling.



supercooled liquid and its stable crystal are equal⁹. For many fragile glass-formers the VTF temperature of structural arrest, T_o , is very close to T_K obtained from calorimetric measurements (typically⁴⁹ $0.9 < T_K/T_o < 1.1$). This again indicates a connection between dynamics and thermodynamics not present at higher temperatures. Equally suggestive is the correspondence between kinetic fragilities based on the temperature dependence of the viscosity (see Fig. 2) and thermodynamic fragilities⁵⁴, based on the temperature dependence of the entropy surplus of the supercooled liquid with respect to its stable crystal.

The energy landscape

A convenient framework for interpreting the complex phenomenology just described is provided by the energy landscape⁴⁴. This is the name generally given to the potential energy function of an N -body system $\Phi(r_1, \dots, r_N)$, where the vectors r_i comprise position, orientation and vibration coordinates. In condensed phases, whether liquid or solid, every molecule experiences simultaneous interactions with numerous neighbours. Under these conditions it is convenient to consider the full N -body Φ -function. The landscape is a multidimensional surface. For the simplest case of N structureless particles possessing no internal orientational and vibrational degrees of freedom, the landscape is a $(3N + 1)$ -dimensional object. Figure 5 is a schematic illustration of an energy landscape. The quantities of interest are the number of potential energy minima (also called inherent structures) of a given depth (Box 2), and the nature of the saddle points separating neighbouring minima. More than 30 years ago, Goldstein articulated a topographic viewpoint of condensed phases⁵⁵ that has come to be known as the energy landscape paradigm. His seminal ideas have since been applied to protein folding^{56–64}, the mechanical properties of glasses^{65–67}, shear-enhanced diffusion⁶⁸ and the dynamics of supercooled liquids^{69–71}.

Landscape sampling

For an N -body material system in a volume V , the landscape is fixed. The manner in which a material system samples its landscape as a

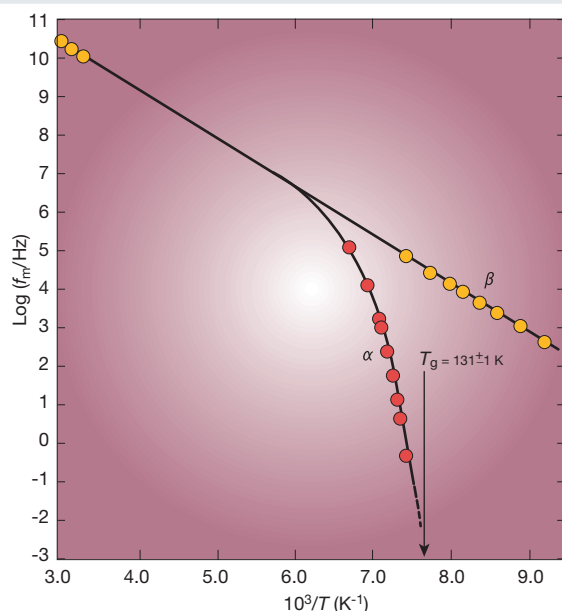


Figure 3 Temperature dependence of the peak dielectric relaxation frequency of the glass-forming mixture chlorobenzene/*cis*-decalin (molar ratio 17.2/82.8%). At high enough temperature there is a single relaxation mechanism. In the moderately supercooled regime the peak splits into slow (α) and fast (β) relaxations, of which α -processes exhibit non-Arrhenius temperature dependence and vanish at T_g . (Adapted from refs 9 and 41.)

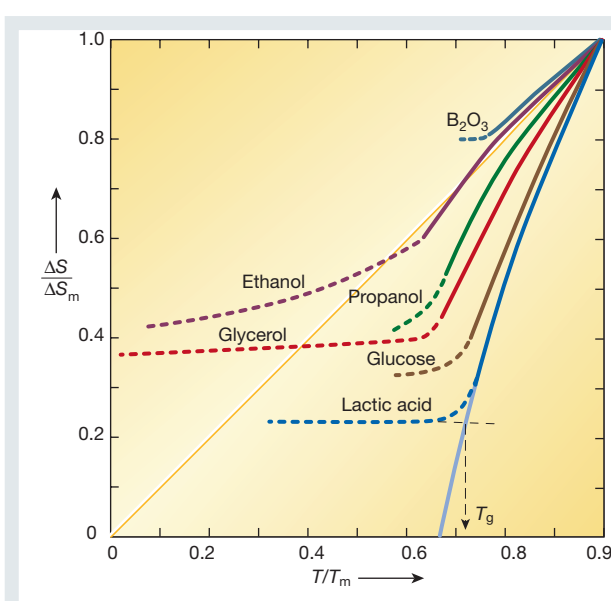


Figure 4 Temperature dependence of the entropy difference between several supercooled liquids and their stable crystals at atmospheric pressure. ΔS_m is the melting entropy and T_m is the melting temperature. The glass transition always intervenes before the vanishing of the entropy surplus. For fragile liquids such as lactic acid, however, an entropy crisis is imminent. (Adapted from ref. 45.)

function of temperature provides information on its dynamic behaviour⁷⁰. The way that a landscape deforms as a result of changes in density provides information on the mechanical properties of a material system⁷². Figure 6 shows the average inherent structure energy for a mixture of unequal-sized atoms, as a function of the temperature of the equilibrated liquid^{70,73}. In these calculations, molecular dynamics simulations of the binary mixture were performed to generate configurations. Periodically, the system's energy was minimized, yielding mechanically stable inherent structures, the average energy of which is reported in the figure. At high temperatures the inherent structure energy is virtually temperature-independent, and appears to have reached a plateau. When the system has sufficient kinetic energy to sample its entire energy landscape, the overwhelming number of minima that it samples are shallow, reflecting the fact that deep minima are very rare (Box 2). But as the reduced temperature decreases below about $T = 1$, the system is unable to surmount the highest energy barriers, and is therefore forced to sample the much rarer deeper minima (Box 2). When this happens, the kinetics of structural relaxation changes from exponential to stretched exponential, and the activation energy (and entropy) associated with structural relaxation become super-Arrhenius, that is to say they increase with decreasing temperature⁷⁰.

These calculations established a connection between changes in dynamics and the manner in which the static and thermodynamic energy landscape is sampled as a function of temperature. Figure 6 also shows that at a low enough temperature the system becomes stuck in a single minimum, the depth of which increases as the cooling rate decreases. This corresponds to the glass transition. Another important observation of this study was the existence of a temperature $T \approx 0.45$, below which the height of the barriers separating sampled inherent structures increases abruptly. This temperature was found to correspond closely to the crossover temperature predicted by mode-coupling theory (MCT; see below) for this system. Here again, it is the manner in which the system samples its landscape, not the landscape itself, that changes with temperature. (See ref. 74 for a recent, different interpretation of landscape sampling at this temperature.)

The landscape picture provides a natural separation of low-temperature molecular motion into sampling distinct potential

energy minima, and vibration within a minimum. It is possible to separate formally the corresponding configurational and vibrational contributions to a liquid's properties^{75,76}. In two important computational studies, the configurational entropy was calculated by probing systematically the statistics governing the sampling of potential energy minima^{77,78} (Box 2). Using this technique, a remarkable connection between configurational entropy and diffusion was identified in liquid water⁷⁹. One of water's distinguishing anomalies is the fact that, at sufficiently low temperature, its diffusivity increases upon compression⁸⁰. As shown in Fig. 7, diffusivity maxima are correlated strongly with configurational entropy maxima, the respective loci coinciding within numerical error.

The results shown in Fig. 7 and the success of the Adam–Gibbs equation in describing experimental data on relaxation in a wide variety of systems⁵² indicate that there exists a scaling relationship between the depth distribution of basins and the height of the saddle points along paths connecting neighbouring basins. Such scaling is not a mathematical necessity, but arises from the nature of real molecular interactions. The topographic nature of this statistical scaling relationship between minima and saddle points is poorly understood (but see the recent computational investigation of saddle points⁷⁴). Its elucidation will explain the origin of the connection between the dynamics and thermodynamics of glass-forming liquids, and constitutes the principal theoretical challenge in this field.

Strong versus fragile behaviour

The extent to which the shear viscosity η deviates from Arrhenius behaviour, $\eta = \eta_0 \exp(E/k_B T)$, constitutes the basis of the classification of liquids as either strong or fragile (Fig. 2). Molten SiO₂, often considered as the prototypical strong glass-former, displays an almost constant activation energy of 180 kcal mol⁻¹ (ref. 81). This constancy indicates that the underlying mechanism, presumably breaking and reformation of Si–O bonds, applies throughout the entire landscape⁴. In contrast, the viscosity of OTP — the canonical fragile glass-former — deviates markedly from Arrhenius behaviour⁸², showing an effective activation energy ($d \ln \eta / d(1/T)$) that increases 20-fold, from one-quarter of the heat of vaporization for the liquid above its melting point to roughly five times the heat of vaporization near T_g . This means that OTP's landscape is very heterogeneous. The basins sampled at high temperature allow relaxation by surmounting low barriers involving the rearrangement of a small number of molecules. The very large activation energy at $T \approx T_g$, on the other hand, corresponds to the cooperative rearrangement of many molecules. These differences between strong and fragile behaviour imply a corresponding topographic distinction between the two

archetypal landscapes. Aside from multiplicity due to permutational symmetry, strong landscapes may consist of a single 'megabasin', whereas fragile ones display a proliferation of well-separated 'megabasins' (Fig. 8).

Cooperative rearrangements such as those that must occur in OTP are unlikely to consist of elementary transitions between adjacent basins. Rather, the likely scenario involves a complicated sequence of elementary transitions. At low temperatures, these rearrangements should be rare and long-lived on the molecular timescale. Furthermore, the diversity of deep landscape traps and of the pathways of configuration space that connect them should result in a broad spectrum of relaxation times, as required for the stretched exponential function in equation (2). This in turn suggests that supercooled fragile liquids are dynamically heterogeneous, probably consisting at any instant of mostly non-diffusing molecules with a few 'hot spots' of mobile molecules. This dynamic heterogeneity³⁹ has both experimental^{29,30,36} and computational^{31–35} support.

The inverse relation between the self-diffusion coefficient and viscosity embodied in the Stokes–Einstein equation is based on macroscopic hydrodynamics that treats the liquid as a continuum.

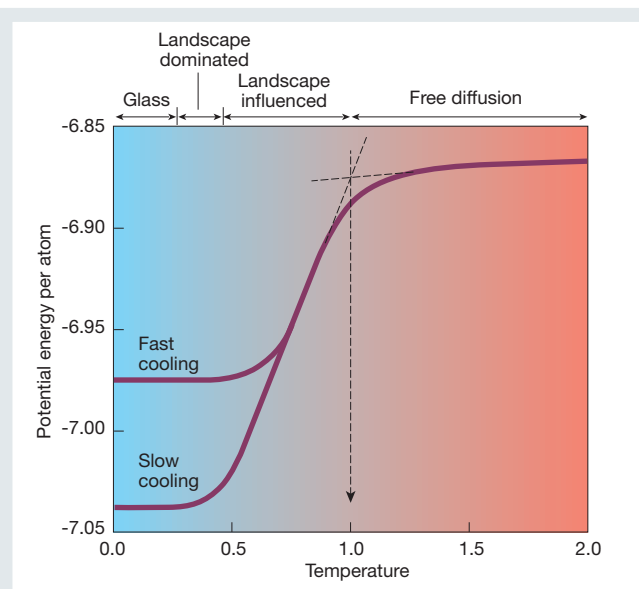


Figure 6 Mean inherent structure energy per particle of a binary mixture of unequal-sized Lennard–Jones atoms, as a function of the temperature of the equilibrated liquid from which the inherent structures were generated by energy minimization. Molecular dynamics simulations at constant energy and density were performed over a range of temperatures for 256 Lennard–Jones atoms, of which 20% are of type A and 80% are of type B. The Lennard–Jones size and energy parameters are $\sigma_{AA} = 1$, $\sigma_{BB} = 0.88$, $\sigma_{AB} = 0.8$, and $\epsilon_{AA} = 1$, $\epsilon_{BB} = 0.5$, $\epsilon_{AB} = 1.5$, respectively. Length, temperature, energy and time are expressed in units of σ_{AA} , ϵ_{AA}/k_B , ϵ_{AA} and $\sigma_{AA}(m/\epsilon_{AA})^{1/2}$, respectively, with m representing the mass of the particles. Simulations were performed at a density of 1.2. The fast and slow cooling rates are 1.08×10^{-3} and 3.33×10^{-6} . When $T > 1$, the system has sufficient kinetic energy to sample the entire energy landscape, and the overwhelming number of sampled energy minima are shallow. Under these conditions, the system exhibits a temperature-independent activation energy for structural relaxation (calculations not shown). Between $T = 1$ and $T \approx 0.45$, the activation energy increases upon cooling, the dynamics become 'landscape-influenced', and the mechanically stable configurations sampled are strongly temperature-dependent. Below $T \approx 0.45$, the height of the barriers separating sampled adjacent energy minima seems to increase abruptly (calculations not shown). This is the 'landscape-dominated' regime. In it, particles execute rare jumps over distances roughly equal to interparticle separations. The crossover between landscape-influenced and landscape-dominated behaviour corresponds closely with the mode-coupling transition temperature^{70,82}. (Adapted from refs 70 and 72.)

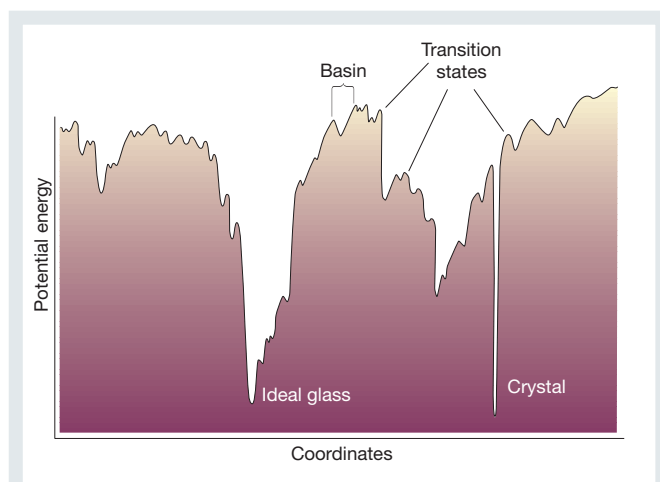


Figure 5 Schematic illustration of an energy landscape. The x-axis represents all configurational coordinates. (Adapted from ref. 44.)

This picture must clearly break down in supercooled fragile liquids, which are dynamically heterogeneous. The failure of the Stokes–Einstein equation, referred to above as one of the distinguishing characteristics of fragile supercooled liquids, is therefore qualitatively understandable. Plausible models for the low-temperature enhancement of diffusive motion relative to hydrodynamic expectations based on the viscosity have been proposed^{83–85}, but an accurate predictive theory is missing. The landscape viewpoint also provides a plausible interpretation for the α/β -relaxation decoupling shown in Fig. 3 — α -relaxations correspond to configurational sampling of neighbouring megabins (Fig. 8), whereas β -processes are thought to correspond to elementary relaxations between contiguous basins⁴⁴. Direct computational evidence of this interpretation is not available.

Avoided singularities

Alternative viewpoints to the landscape perspective have also contributed to current understanding of some aspects of supercooling and the glass transition. Two such interpretations invoke a narrowly avoided singularity above T_g .

According to MCT⁸⁶, structural arrest occurs as a result of the following feedback mechanism: (i) shear-stress relaxation occurs primarily through diffusive motion; (ii) diffusion and viscosity are inversely related; and (iii) viscosity is proportional to shear-stress relaxation time. These facts lead to a viscosity feedback whereby

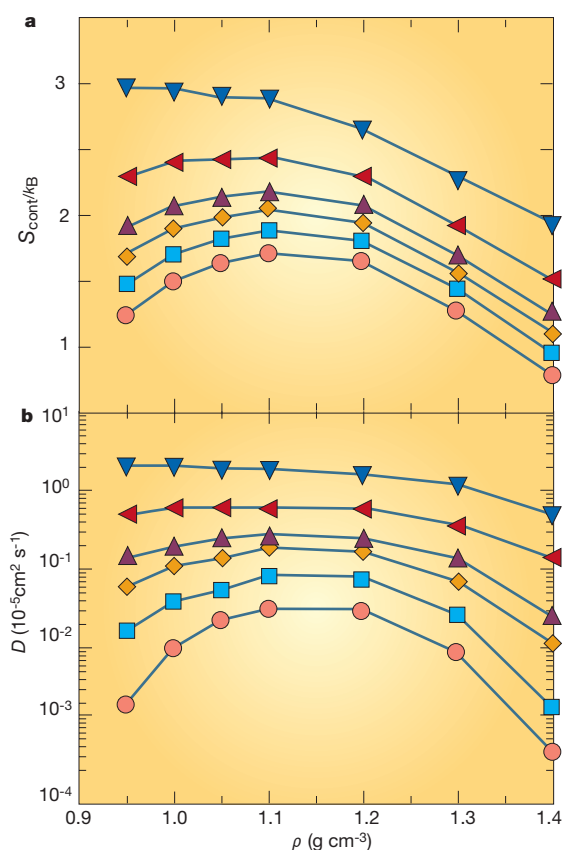
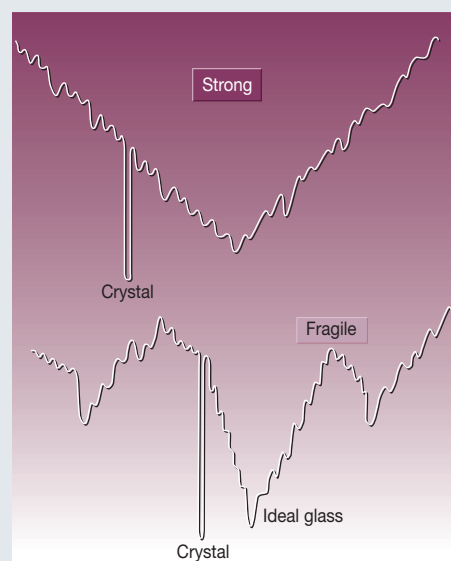


Figure 7 Relationship between diffusivity (D) and configurational entropy (S_{conf}) of supercooled water⁷⁹ at six different temperatures. Filled and open symbols from top to bottom represent the following temperatures: 300 K, 260 K, 240 K, 230 K, 220 K and 210 K; ρ is density. The configurational entropy, which is related to the number of potential energy minima of a given depth (Box 2), was calculated by subtracting the vibrational contribution from the total entropy. The calculations involved performing molecular dynamics simulations of the extended simple point charge (SPC/E) model of water¹⁰² over a range of temperatures and densities. (Adapted from ref. 79.)

Figure 8 Schematic representation of the energy landscapes of strong and fragile substances. The potential energy increases vertically, and the horizontal direction represents collective configurational coordinates.



structural arrest occurs as a purely dynamic singularity, that is to say it is not accompanied by thermodynamic signatures such as a diverging correlation length. What is now known as the idealized MCT^{87,88} predicts structural arrest to occur at a temperature T_x . Initially, therefore, it was thought that MCT was a useful theory for the laboratory-generated glass transition. It is now widely understood that this is not the case, as one finds that $T_x > T_g$, and the MCT-predicted singularity does not occur. In subsequent modifications of the theory⁸⁹, additional relaxation mechanisms occur, often referred to as ‘hopping’ or activated motions, which restore ergodicity (the system’s ability to sample all configurations) below T_x , thereby avoiding a kinetic singularity. These additional relaxation modes arise as a result of a coupling between fluctuations in density and momentum.

Although not a theory of the glass transition, MCT accurately describes many important aspects of relaxation dynamics in liquids above or moderately below their melting temperatures. In particular, the theory makes detailed predictions about the behaviour of the intermediate scattering function F , an experimentally observable quantity that measures the decay of density fluctuations. After a fast initial decay due to microscopic intermolecular collisions, MCT predicts that the decay of F obeys the following sequence (Fig. 9): (i) power-law decay towards a plateau, according to $F = f + At^{-a}$; (ii) a second power-law decay away from the plateau value $F = f - Bt^b$; and (iii) slow relaxation at longer times, which can be fitted by the KWW function $F = \exp[-(t/\tau)^\beta]$. Here, f is the plateau value of the scattering function, which only appears at sufficiently low temperature; t is time; A , B , a and b are constants; τ is the characteristic, temperature-dependent relaxation time; and $\beta < 1$ is the KWW stretch exponent. The basic accuracy of these detailed predictions has been verified experimentally and in computer simulations^{90–92}.

Kivelson and co-workers have proposed a theory of supercooled liquids that is based also on an avoided singularity^{24,93–95}. According to this viewpoint, the liquid has an energetically preferred local structure that differs from the structure in the actual crystalline phase. The system is prevented from crystallizing into a reference crystal with the preferred local structure because of geometric frustration owing to the fact that the latter does not tile space. An example of such energetically favoured but non-space-tiling local structure is the icosahedral packing seen in computer simulations of the supercooled Lennard–Jones liquid⁷³. At a temperature T^* the system would, but for frustration, crystallize into the reference crystal. Instead, strain build-up causes the system to break up into frustration-limited domains, thereby avoiding a phase transition (singularity) at T^* . The avoided transition temperature T^* acts as a critical point, below which two length scales emerge, both of which are large compared to

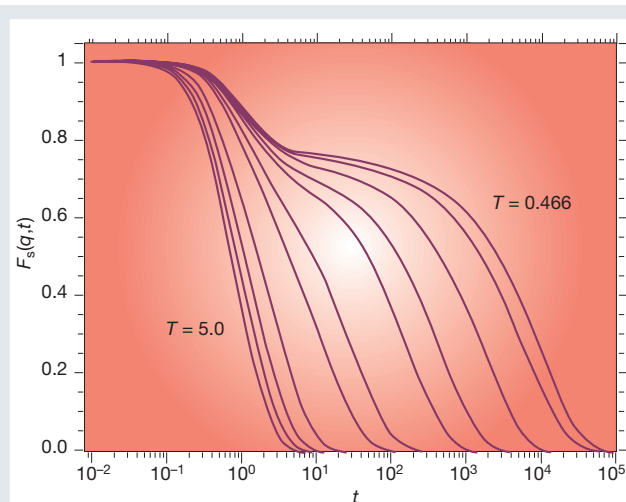


Figure 9 Evolution of the self-intermediate scattering function for A-type atoms for the same supercooled Lennard–Jones mixture as in Fig. 6, at $q\sigma_{AA} = 7.251$, corresponding to the first peak of the static structure factor of species A (ref. 92). Here q is the magnitude of the wave vector. Temperature and time are in units of ϵ_{AA}/k_B and $\sigma_{AA}(m/48\epsilon_{AA})^{1/2}$, respectively. Temperatures from left to right are 5, 4, 3, 2, 1, 0.8, 0.6, 0.55, 0.5, 0.475, and 0.466. The self-intermediate scattering function is the space Fourier transform of the van Hove function $G_s(r, t)$, which is proportional to the probability of observing a particle at $r \pm dr$ at time t given that the same particle was at the origin at $t = 0$. Note the two-step relaxation behaviour upon decreasing T . Molecular dynamics simulations of 1,000 atoms. (Adapted from refs 9 and 92.)

molecular dimensions. One is the critical correlation length, which governs density fluctuations in the absence of frustration. The second is the frustration-limited domain size. From these considerations there emerge predictions on the temperature dependence of the viscosity. Experimental data analysed according to the theory display universality²⁴, but at the expense of introducing a number of fitting parameters. The improvement with respect to competing interpretations is a matter of controversy²⁵.

Challenges and open questions

Important aspects of the complex behaviour of viscous liquids close to the glass transition can be explained qualitatively from the energy landscape perspective. Making this descriptive picture quantitative and predictive is a major challenge. This will require investigating how basic landscape features such as the basin enumeration function depend on molecular architecture and, for a given substance or mixture, on density (see ref. 96 for an example of such a calculation). Equally important is the translation of qualitative pictures such as Fig. 8 into precise measures of strength and fragility based on the basin enumeration function. Uncovering the topographic nature of the scaling relationship between basin minima and saddle points holds the key to understanding the relationship between kinetics and thermodynamics in deeply supercooled liquids. All of these calculations are in principle straightforward, but computationally at the very limit of what is currently feasible. The development of theoretical models⁹⁷ is therefore of paramount importance.

MCT and the landscape perspective offer complementary viewpoints of the same phenomena. So far, however, not enough effort has been devoted to bridging the gap that separates these two approaches. Recent calculations^{70,74} offer the promise of establishing a clearer connection between the static landscape viewpoint and the dynamic perspective of MCT. At the least, what is required is a precise landscape-based explanation of what ‘hopping’ and ‘activated processes’ really mean. Additional theoretical viewpoints of supercooling and the glass transition include the instantaneous normal-mode perspective on liquid dynamics⁹⁸ and thermodynamic

treatments of the vitreous state based on invoking analogies to spin glasses^{99–101}. Establishing a coherent theoretical perspective on supercooled liquids and glasses is important. We believe that the landscape formalism offers the natural technical tools for accomplishing this task.

- Angell, C. A. Formation of glasses from liquids and biopolymers. *Science* **267**, 1924–1935 (1995).
- Blanshard, J. M. V. & Lillford, P. (eds) *The Glassy State in Foods* (Nottingham Univ. Press, Nottingham, 1993).
- Crowe, J. H., Carpenter, J. F. & Crowe, L. M. The role of vitrification in anhydrobiosis. *Annu. Rev. Physiol.* **60**, 73–103 (1998).
- Debenedetti, P. G., Stillinger, F. H., Truskett, T. M. & Lewis, C. P. Theory of supercooled liquids and glasses: energy landscape and statistical geometry perspectives. *Adv. Chem. Eng.* (in the press).
- Greer, A. L. Metallic glasses. *Science* **267**, 1947–1953 (1995).
- Jenniskens, P. & Blake, D. F. Structural transitions in amorphous water ice and astrophysical implications. *Science* **265**, 753–756 (1994).
- Anderson, P. W. Through a glass lightly. *Science* **267**, 1615 (1995).
- Angell, C. A., Ngai, K. L., McKenna, G. B., McMillan, P. F. & Martin, S. W. Relaxation in glass-forming liquids and amorphous solids. *J. Appl. Phys.* **88**, 3113–3157 (2000).
- Debenedetti, P. G. *Metastable Liquids. Concepts and Principles* (Princeton Univ. Press, Princeton, 1996).
- Turnbull, D. Under what conditions can a glass be formed? *Contemp. Phys.* **10**, 473–488 (1969).
- Angell, C. A. Structural instability and relaxation in liquid and glassy phases near the fragile liquid limit. *J. Non-Cryst. Solids* **102**, 205–221 (1988).
- Moynihan, C. T. et al. in *The Glass Transition and the Nature of the Glassy State* (eds Goldstein, M. & Simha, R.) *Ann. NY Acad. Sci.* **279**, 15–36 (1976).
- Brüning, R. & Samwer, K. Glass transition on long time scales. *Phys. Rev. B* **46**, 318–322 (1992).
- Ediger, M. D., Angell, C. A. & Nagel, S. R. Supercooled liquids and glasses. *J. Phys. Chem.* **100**, 13200–13212 (1996).
- Vogel, H. Das temperatur-abhängigkeitsgesetz der viskosität von flüssigkeiten. *Phys. Zeit.* **22**, 645–646 (1921).
- Tammann, G. & Hesse, W. Die abhängigkeit der viskosität von der temperatur bei unterkühlten flüssigkeiten. *Z. Anorg. Allg. Chem.* **156**, 245–257 (1926).
- Fulcher, G. S. Analysis of recent measurements of the viscosity of glasses. *J. Am. Ceram. Soc.* **8**, 339 (1925).
- Laughlin, W. T. & Uhlmann, D. R. Viscous flow in simple organic liquids. *J. Phys. Chem.* **76**, 2317–2325 (1972).
- Angell, C. A. in *Relaxations in Complex Systems* (eds Ngai, K. & Wright, G. B.) **1** (Nat'l Technol. Inform. Ser., US Dept. of Commerce, Springfield, VA, 1985).
- Angell, C. A. Relaxation in liquids, polymers and plastic crystals—strong/fragile patterns and problems. *J. Non-Cryst. Solids* **131–133**, 13–31 (1991).
- Green, J. L., Ito, K., Xu, K. & Angell, C. A. Fragility in liquids and polymers: new, simple quantifications and interpretations. *J. Phys. Chem. B* **103**, 3991–3996 (1999).
- Novikov, V. N., Rössler, E., Malinovsky, V. K. & Surovstev, N. V. Strong and fragile liquids in percolation approach to the glass transition. *Europhys. Lett.* **35**, 289–294 (1996).
- Fujimori, H. & Oguni, M. Correlation index $(T_g - T_{\infty})/T_g$ and activation energy ratio $\Delta E_{\infty}/\Delta E_{\text{eff}}$ as parameters characterizing the structure of liquid and glass. *Solid State Commun.* **94**, 157–162 (1995).
- Kivelson, D., Tarjus, G., Zhao, X. & Kivelson, S. A. Fitting of viscosity: distinguishing the temperature dependencies predicted by various models of supercooled liquids. *Phys. Rev. E* **53**, 751–758 (1996).
- Cummins, H. Z. Comment on “Fitting of viscosity: distinguishing the temperature dependencies predicted by various models of supercooled liquids”. *Phys. Rev. E* **54**, 5870–5872 (1996).
- Kohlrausch, R. Theorie des elektrischen rückstandes in der leidener flasche. *Ann. Phys. Chem. (Leipzig)* **91**, 179–214 (1874).
- Williams, G. & Watts, D. C. Non-symmetrical dielectric relaxation behavior arising from a simple empirical decay function. *Trans. Faraday Soc.* **66**, 80–85 (1970).
- Richert, R. & Blumen, A. in *Disorder Effects on Relaxational Processes* (eds Richert, R. & Blumen, A.) **1–7** (Springer, Berlin, 1994).
- Cicerone, M. T. & Ediger, M. D. Relaxation of spatially heterogeneous dynamic domains in supercooled ortho-terphenyl. *J. Chem. Phys.* **103**, 5684–5692 (1995).
- Cicerone, M. T. & Ediger, M. D. Enhanced translation of probe molecules in supercooled o-terphenyl: signature of spatially heterogeneous dynamics? *J. Chem. Phys.* **104**, 7210–7218 (1996).
- Mečuk, A. I., Ramos, R. A., Gould, H., Klein, W. & Mountain, R. D. Long-lived structures in fragile glass-forming liquids. *Phys. Rev. Lett.* **75**, 2522–2525 (1995).
- Hurley, M. M. & Harrowell, P. Non-gaussian behavior and the dynamical complexity of particle motion in a dense two-dimensional liquid. *J. Chem. Phys.* **105**, 10521–10526 (1996).
- Perera, D. N. & Harrowell, P. Measuring diffusion in supercooled liquids: the effect of kinetic inhomogeneities. *J. Chem. Phys.* **104**, 2369–2375 (1996).
- Perera, D. N. & Harrowell, P. Consequence of kinetic inhomogeneities in glasses. *Phys. Rev. E* **54**, 1652–1662 (1996).
- Donati, C., Glotzer, S. C., Poole, P. H., Kob, W. & Plimpton, S. J. Spatial correlations of mobility and immobility in a glass-forming Lennard–Jones liquid. *Phys. Rev. E* **60**, 3107–3119 (1999).
- Böhmer, R., Hinze, G., Diezemann, G., Geil, B. & Sillescu, H. Dynamic heterogeneity on supercooled ortho-terphenyl studied by multidimensional deuterium NMR. *Europhys. Lett.* **36**, 55–60 (1996).
- Wang, C.-Y. & Ediger, M. D. How long do regions of different dynamics persist in supercooled o-terphenyl? *J. Phys. Chem. B* **103**, 4177–4184 (1999).
- Vidal Russell, E. & Israeloff, N. E. Direct observation of molecular cooperativity near the glass transition. *Nature* **408**, 695–698 (2000).
- Ediger, M. D. Spatially heterogeneous dynamics in supercooled liquids. *Annu. Rev. Phys. Chem.* **51**, 99–128 (2000).
- Fujara, F., Geil, B., Sillescu, H. H. & Fleischer, G. Translational and rotational diffusion in supercooled ortho-terphenyl close to the glass transition. *Z. Phys. B Cond. Matt.* **88**, 195–204 (1992).
- Johari, G. P. Intrinsic mobility of molecular glasses. *J. Chem. Phys.* **58**, 1766–1770 (1973).
- Johari, G. P. & Goldstein, M. Viscous liquids and the glass transition. II. Secondary relaxations in glasses of rigid molecules. *J. Chem. Phys.* **53**, 2372–2388 (1970).

43. Rössler, E., Warschewski, U., Eiermann, P., Sokolov, A. P. & Quitmann, D. Indications for a change of transport mechanism in supercooled liquids and the dynamics close and below T_g . *J. Non-Cryst. Solids* **172–174**, 113–125 (1994).
44. Stillinger, F. H. A topographic view of supercooled liquids and glass formation. *Science* **267**, 1935–1939 (1995).
45. Kauzmann, W. The nature of the glassy state and the behavior of liquids at low temperatures. *Chem. Rev.* **43**, 219–256 (1948).
46. Simon, F. Über den Zustand der unterkühlten Flüssigkeiten und Gläser. *Z. Anorg. Allg. Chem.* **203**, 219–227 (1931).
47. Wolynes, P. G. Aperiodic crystals: biology, chemistry and physics in a fugue with stretto. *AIP Conf. Proc.* **180**, 39–65 (1988).
48. Wolynes, P. G. Entropy crises in glasses and random heteropolymers. *J. Res. Natl. Inst. Standards Technol.* **102**, 187–194 (1997).
49. Angell, C. A. Landscapes with megabasins: polyamorphism in liquids and biopolymers and the role of nucleation in folding and folding diseases. *Physica D* **107**, 122–142 (1997).
50. Gibbs, J. H. & DiMarzio, E. A. Nature of the glass transition and the glassy state. *J. Chem. Phys.* **28**, 373–383 (1958).
51. Adam, G. & Gibbs, J. H. On the temperature dependence of cooperative relaxation properties in glass-forming liquids. *J. Chem. Phys.* **43**, 139–146 (1965).
52. Richert, R. & Angell, C. A. Dynamics of glass-forming liquids. V. On the link between molecular dynamics and configurational entropy. *J. Chem. Phys.* **108**, 9016–9026 (1998).
53. Williams, M. L., Landel, R. F. & Ferry, J. D. The temperature dependence of the relaxation mechanisms in amorphous polymers and other glass-forming liquids. *J. Am. Chem. Soc.* **77**, 3701–3707 (1955).
54. Ito, K., Moynihan, C. T. & Angell, C. A. Thermodynamic determination of fragility in liquids and a fragile-to-strong liquid transition in water. *Nature* **398**, 492–495 (1999).
55. Goldstein, M. Viscous liquids and the glass transition: a potential energy barrier picture. *J. Chem. Phys.* **51**, 3728–3739 (1969).
56. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).
57. Nienhaus, G. U., Müller, J. D., McMahon, B. H. & Frauenfelder, H. Exploring the conformational energy landscape of proteins. *Physica D* **107**, 297–311 (1997).
58. Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. Free energy landscape for protein folding kinetics: intermediates, traps, and multiple pathways in theory and lattice model simulations. *J. Chem. Phys.* **101**, 6052–6062 (1994).
59. Saven, J. G., Wang, J. & Wolynes, P. G. Kinetics of protein folding: the dynamics of globally connected rough energy landscapes with biases. *J. Chem. Phys.* **101**, 11037–11043 (1994).
60. Wang, J., Onuchic, J. & Wolynes, P. Statistics of kinetic pathways on biased rough energy landscapes with applications to protein folding. *Phys. Rev. Lett.* **76**, 4861–4864 (1996).
61. Plotkin, S. S., Wang, J. & Wolynes, P. G. Correlated energy landscape model for finite, random heteropolymers. *Phys. Rev. E* **53**, 6271–6296 (1996).
62. Becker, O. M. & Karplus, M. The topology of multidimensional potential energy surfaces: theory and application to peptide structure and kinetics. *J. Chem. Phys.* **106**, 1495–1517 (1997).
63. Dill, K. A. & Chan, H. S. From Levinthal to pathways and funnels. *Nature Struct. Biol.* **4**, 10–19 (1997).
64. Klepeis, J. L., Floudas, C. A., Morikis, D. & Lambris, J. D. Predicting peptide structure using NMR data and deterministic global optimization. *J. Comp. Chem.* **20**, 1354–1370 (1999).
65. Lacks, D. J. Localized mechanical instabilities and structural transformations in silica glass under high pressure. *Phys. Rev. Lett.* **80**, 5385–5388 (1998).
66. Malandro, D. L. & Lacks, D. J. Volume dependence of potential energy landscapes in glasses. *J. Chem. Phys.* **107**, 5804–5810 (1997).
67. Malandro, D. L. & Lacks, D. J. Relationships of shear-induced changes in the potential energy landscape to the mechanical properties of ductile glasses. *J. Chem. Phys.* **110**, 4593–4601 (1999).
68. Malandro, D. L. & Lacks, D. J. Molecular-level instabilities and enhanced self-diffusion in flowing liquids. *Phys. Rev. Lett.* **81**, 5576–5579 (1998).
69. Schulz, M. Energy landscape, minimum points, and non-Arrhenius behavior of supercooled liquids. *Phys. Rev. B* **57**, 11319–11333 (1998).
70. Sastry, S., Debenedetti, P. G. & Stillinger, F. H. Signatures of distinct dynamical regimes in the energy landscape of a glass-forming liquid. *Nature* **393**, 554–557 (1998).
71. Keyes, T. Dependence of supercooled liquid dynamics on elevation in the energy landscape. *Phys. Rev. E* **59**, 3207–3211 (1999).
72. Debenedetti, P. G., Stillinger, F. H., Truskett, T. M. & Roberts, C. J. The equation of state of an energy landscape. *J. Phys. Chem. B* **103**, 7390–7397 (1999).
73. Jonsson, H. & Andersen, H. C. Icosahedral ordering in the Lennard-Jones crystal and glass. *Phys. Rev. Lett.* **60**, 2295–2298 (1988).
74. Angelani, L., Di Leonardo, R., Ruocco, G., Scala, A. & Sciortino, F. Saddles in the energy landscape probed by supercooled liquids. *Phys. Rev. Lett.* **85**, 5356–5359 (2000).
75. Stillinger, F. H., Debenedetti, P. G. & Sastry, S. Resolving vibrational and structural contributions to isothermal compressibility. *J. Chem. Phys.* **109**, 3983–3988 (1998).
76. Stillinger, F. H. & Debenedetti, P. G. Distinguishing vibrational and structural equilibration contributions to thermal expansion. *J. Phys. Chem. B* **103**, 4052–4059 (1999).
77. Sciortino, F., Kob, W. & Tartaglia, P. Inherent structure entropy of supercooled liquids. *Phys. Rev. Lett.* **83**, 3214–3217 (1999).
78. Büchner, S. & Heuer, A. Potential energy landscape of a model glass former: thermodynamics, anharmonicities, and finite size effects. *Phys. Rev. E* **60**, 6507–6518 (1999).
79. Scala, A., Starr, F. W., La Nave, E., Sciortino, F. & Stanley, H. E. Configurational entropy and diffusivity in supercooled water. *Nature* **406**, 166–169 (2000).
80. Prielmeier, F. X., Lang, E. W., Speedy, R. J. & Lüdemann, H.-D. Diffusion in supercooled water to 300 Mpa. *Phys. Rev. Lett.* **59**, 1128–1131 (1987).
81. Mackenzie, J. D. Viscosity-temperature relationship for network liquids. *J. Am. Ceram. Soc.* **44**, 598–601 (1961).
82. Greet, R. J. & Turnbull, D. Glass transition in o-terphenyl. *J. Chem. Phys.* **46**, 1243–1251 (1967).
83. Stillinger, F. H. & Hodgdon, J. A. Translation-rotation paradox for diffusion in fragile glass-forming liquids. *Phys. Rev. E* **50**, 2064–2068 (1994).
84. Tarjus, G. & Kivelson, D. Breakdown of the Stokes-Einstein relation in supercooled liquids. *J. Chem. Phys.* **103**, 3071–3073 (1995).
85. Liu, C. Z.-W. & Openheim, I. Enhanced diffusion upon approaching the kinetic glass transition. *Phys. Rev. E* **53**, 799–802 (1996).
86. Gesztzi, T. Pre-vitrification by viscosity feedback. *J. Phys. C* **16**, 5805–5814 (1983).
87. Bengtzelius, U., Götz, W. & Sjölander, A. Dynamics of supercooled liquids and the glass transition. *J. Phys. C* **17**, 5915–5934 (1984).
88. Götz, W. & Sjögren, L. Relaxation processes in supercooled liquids. *Rep. Prog. Phys.* **55**, 241–376 (1992).
89. Götz, W. & Sjögren, L. The mode coupling theory of structural relaxations. *Transp. Theory Stat. Phys.* **24**, 801–853 (1995).
90. Götz, W. Recent tests of the mode-coupling theory for glassy dynamics. *J. Phys. Cond. Matt.* **11**, A1–A45 (1999).
91. Kob, W. Computer simulations of supercooled liquids and glasses. *J. Phys. Cond. Matt.* **11**, R85–R115 (1999).
92. Kob, W. & Andersen, H. C. Testing mode-coupling theory for a supercooled binary Lennard-Jones mixture: the van Hove correlation function. *Phys. Rev. E* **51**, 4626–4641 (1995).
93. Kivelson, D., Kivelson, S. A., Zhao, X., Nussinov, Z. & Tarjus, G. A thermodynamic theory of supercooled liquids. *Physica A* **219**, 27–38 (1995).
94. Kivelson, D. & Tarjus, G. SuperArrhenius character of supercooled glass-forming liquids. *J. Non-Cryst. Solids* **235–237**, 86–100 (1998).
95. Kivelson, D. & Tarjus, G. The Kauzmann paradox interpreted via the theory of frustration-limited domains. *J. Chem. Phys.* **109**, 5481–5486 (1998).
96. Sastry, S. The relationship between fragility, configurational entropy and the potential energy landscape of glass-forming liquids. *Nature* **409**, 164–167 (2001).
97. Speedy, R. J. Relations between a liquid and its glasses. *J. Phys. Chem. B* **103**, 4060–4065 (1999).
98. Keyes, T. Instantaneous normal mode approach to liquid state dynamics. *J. Phys. Chem. A* **101**, 2921–2930 (1997).
99. Kirkpatrick, T. R. & Wolynes, P. G. Stable and metastable states in mean-field Potts and structural glasses. *Phys. Rev. B* **36**, 8552–8564 (1987).
100. Kirkpatrick, T. R., Thirumalai, D. & Wolynes, P. G. Scaling concepts for the dynamics of viscous liquids near an ideal glassy state. *Phys. Rev. A* **40**, 1045–1054 (1989).
101. Mézard, M. & Parisi, G. Thermodynamics of glasses: a first principles computation. *Phys. Rev. Lett.* **82**, 747–750 (1999).
102. Berendsen, H. J., Grigera, J. R. & Stroatsma, T. P. The missing term in effective pair potentials. *J. Phys. Chem.* **91**, 6269–6271 (1987).
103. Stillinger, F. H. Supercooled liquids, glass transitions, and the Kauzmann paradox. *J. Chem. Phys.* **88**, 7818–7825 (1988).
104. Santen, L. & Krauth, W. Absence of thermodynamic phase transition in a model glass former. *Nature* **405**, 550–551 (2000).
105. Wilks, J. *The Properties of Liquid and Solid Helium* (Clarendon, Oxford, 1967).
106. Rastogi, S., Höhne, G. W. H. & Keller, A. Unusual pressure-induced phase behavior in crystalline Poly(4-methylpentene-1): calorimetric and spectroscopic results and further implications. *Macromolecules* **32**, 8897–8909 (1999).
107. Greer, A. L. Too hot to melt. *Nature* **404**, 134–135 (2000).
108. Stillinger, F. H. Exponential multiplicity of inherent structures. *Phys. Rev. E* **59**, 48–51 (1999).
109. Stillinger, F. H. Enumeration of isobaric inherent structures for the fragile glass former o-terphenyl. *J. Phys. Chem. B* **102**, 2807–2810 (1998).

Acknowledgements

P.G.D.'s work is supported by the US Department of Energy.

Exploring complex networks

Steven H. Strogatz

Department of Theoretical and Applied Mechanics and Center for Applied Mathematics, 212 Kimball Hall, Cornell University, Ithaca, New York 14853-1503, USA (e-mail: strogatz@cornell.edu)

The study of networks pervades all of science, from neurobiology to statistical physics. The most basic issues are structural: how does one characterize the wiring diagram of a food web or the Internet or the metabolic network of the bacterium *Escherichia coli*? Are there any unifying principles underlying their topology? From the perspective of nonlinear dynamics, we would also like to understand how an enormous network of interacting dynamical systems — be they neurons, power stations or lasers — will behave collectively, given their individual dynamics and coupling architecture. Researchers are only now beginning to unravel the structure and dynamics of complex networks.

Networks are on our minds nowadays. Sometimes we fear their power — and with good reason. On 10 August 1996, a fault in two power lines in Oregon led, through a cascading series of failures, to blackouts in 11 US states and two Canadian provinces, leaving about 7 million customers without power for up to 16 hours¹. The Love Bug worm, the worst computer attack to date, spread over the Internet on 4 May 2000 and inflicted billions of dollars of damage worldwide.

In our lighter moments we play parlour games about connectivity. 'Six degrees of Marlon Brando' broke out as a nationwide fad in Germany, as readers of *Die Zeit* tried to connect a falafel vendor in Berlin with his favourite actor through the shortest possible chain of acquaintances². And during the height of the Lewinsky scandal, the *New York Times* printed a diagram³ of the famous people within 'six degrees of Monica'.

Meanwhile scientists have been thinking about networks too. Empirical studies have shed light on the topology of food webs^{4,5}, electrical power grids, cellular and metabolic networks^{6–9}, the World-Wide Web¹⁰, the Internet backbone¹¹, the neural network of the nematode worm *Caenorhabditis elegans*¹², telephone call graphs¹³, coauthorship and citation networks of scientists^{14–16}, and the quintessential 'old-boy' network, the overlapping boards of directors of the largest companies in the United States¹⁷ (Fig. 1). These databases are now easily accessible, courtesy of the Internet. Moreover, the availability of powerful computers has made it feasible to probe their structure; until recently, computations involving million-node networks would have been impossible without specialized facilities.

Why is network anatomy so important to characterize? Because structure always affects function. For instance, the topology of social networks affects the spread of information and disease, and the topology of the power grid affects the robustness and stability of power transmission.

From this perspective, the current interest in networks is part of a broader movement towards research on complex systems. In the words of E. O. Wilson¹⁸, "The greatest challenge today, not just in cell biology and ecology but in all of science, is the accurate and complete description of complex systems. Scientists have broken down many kinds of systems. They think they know most of the elements and forces. The next task is to reassemble them, at least in mathematical models that capture the key properties of the entire ensembles."

But networks are inherently difficult to understand, as the following list of possible complications illustrates.

1. Structural complexity: the wiring diagram could be an intricate tangle (Fig. 1).
2. Network evolution: the wiring diagram could change over time. On the World-Wide Web, pages and links are created and lost every minute.
3. Connection diversity: the links between nodes could have different weights, directions and signs. Synapses in

Box 1

Nonlinear dynamics: terminology and concepts⁹⁷

Dynamical systems can often be modelled by differential equations $d\mathbf{x}/dt = \mathbf{v}(\mathbf{x})$, where $\mathbf{x}(t) = (x_1(t), \dots, x_n(t))$ is a vector of state variables, t is time, and $\mathbf{v}(\mathbf{x}) = (v_1(\mathbf{x}), \dots, v_n(\mathbf{x}))$ is a vector of functions that encode the dynamics. For example, in a chemical reaction, the state variables represent concentrations. The differential equations represent the kinetic rate laws, which usually involve nonlinear functions of the concentrations.

Such nonlinear equations are typically impossible to solve analytically, but one can gain qualitative insight by imagining an abstract n -dimensional state space with axes $x_1, \dots, x_n$. As the system evolves, $\mathbf{x}(t)$ flows through state space, guided by the 'velocity' field $d\mathbf{x}/dt = \mathbf{v}(\mathbf{x})$ like a speck carried along in a steady, viscous fluid.

Suppose $\mathbf{x}(t)$ eventually comes to rest at some point $\mathbf{x}^*$. Then the velocity must be zero there, so we call $\mathbf{x}^*$ a fixed point. It corresponds to an equilibrium state of the physical system being modelled. If all small disturbances away from $\mathbf{x}^*$ damp out, $\mathbf{x}^*$ is called a stable fixed point — it acts as an attractor for states in its vicinity.

Another long-term possibility is that $\mathbf{x}(t)$ flows towards a closed loop and eventually circulates around it forever. Such a loop is called a limit cycle. It represents a self-sustained oscillation of the physical system.

A third possibility is that $\mathbf{x}(t)$ might settle onto a strange attractor, a set of states on which it wanders forever, never stopping or repeating. Such erratic, aperiodic motion is considered chaotic if two nearby states flow away from each other exponentially fast. Long-term prediction is impossible in a real chaotic system because of this exponential amplification of small uncertainties or measurement errors

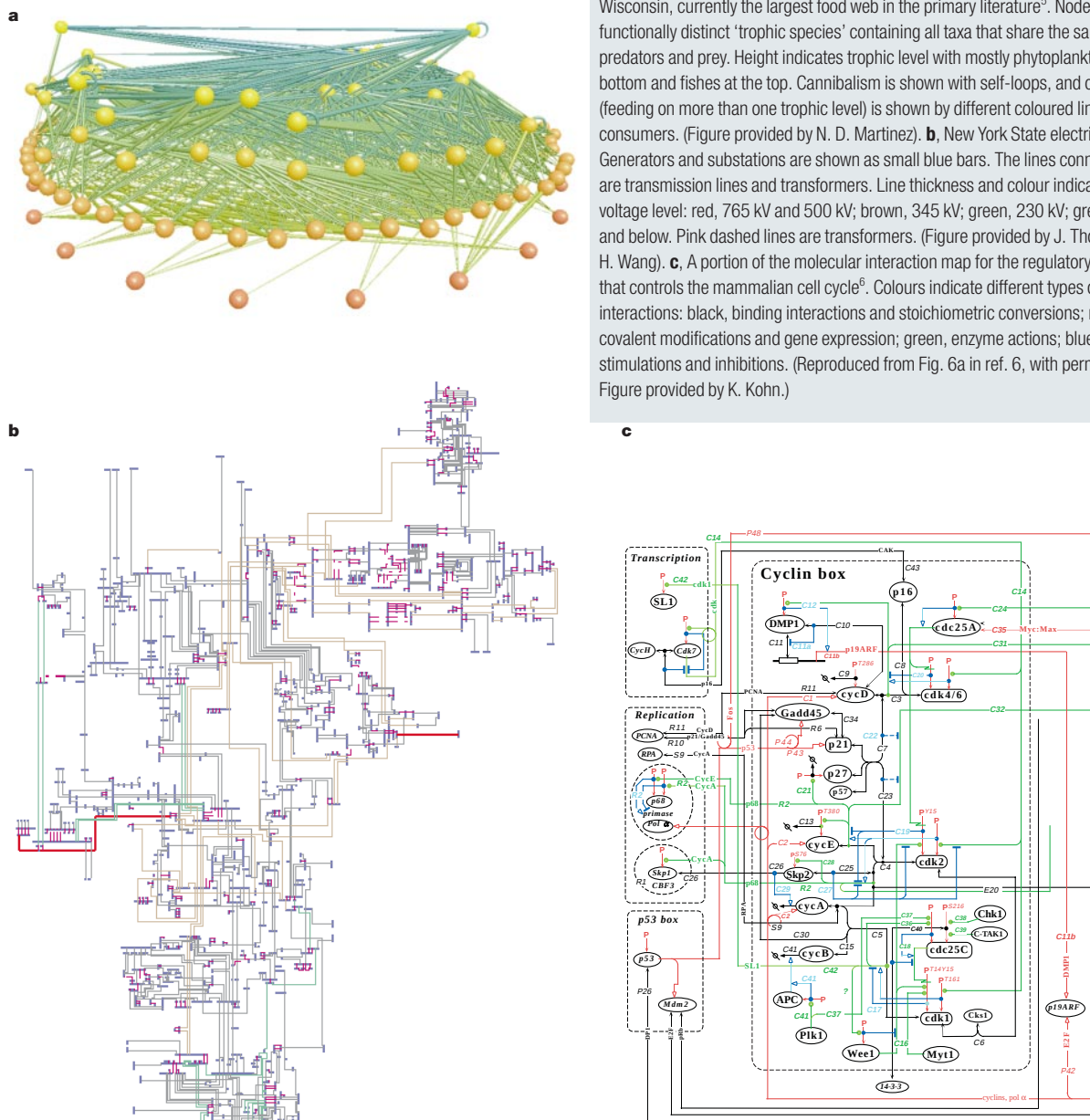
the nervous system can be strong or weak, inhibitory or excitatory.

4. Dynamical complexity: the nodes could be nonlinear dynamical systems. In a gene network or a Josephson junction array, the state of each node can vary in time in complicated ways.
5. Node diversity: there could be many different kinds of nodes. The biochemical network that controls cell division in mammals consists of a bewildering variety of substrates and enzymes⁶, only a few of which are shown in Fig. 1c.
6. Meta-complication: the various complications can influence each other. For example, the present layout of a power grid depends on how it has grown over the years — a case where network evolution (2) affects topology (1). When coupled neurons fire together repeatedly, the connection between them is strengthened; this is the basis of memory and learning. Here nodal dynamics (4) affect connection weights (3).

To make progress, different fields have suppressed certain complications while highlighting others. For instance, in nonlinear dynamics we have tended to favour simple, nearly identical

dynamical systems coupled together in simple, geometrically regular ways. Furthermore we usually assume that the network architecture is static. These simplifications allow us to sidestep any issues of structural complexity and to concentrate instead on the system's potentially formidable dynamics.

Laser arrays provide a concrete example^{19–24}. In the single-mode approximation, each laser is characterized by its time-dependent gain, polarization, and the phase and amplitude of its electric field. These evolve according to four coupled, nonlinear differential equations. We usually hope the laser will settle down to a stable state, corresponding to steady emission of light, but periodic pulsations and even chaotic intensity fluctuations can occur in some cases¹⁹. Now suppose that many identical lasers are arranged side by side in a regular chain²⁰ or ring²¹, interacting with their neighbours by evanescent coupling or by overlap of their electric fields²². Will the lasers lock their phases together spontaneously, or break up into a standing wave pattern, or beat each other into incoherence? From a technological standpoint, self-synchronization would be the most desirable outcome, because a perfectly coherent array of N lasers would



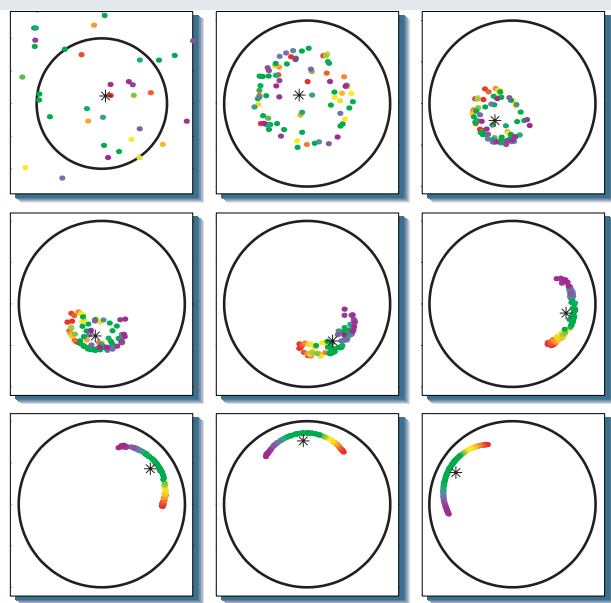


Figure 2 Spontaneous synchronization in a network of limit-cycle oscillators with distributed natural frequencies. The state of each oscillator is represented geometrically as a dot in the complex plane. The amplitude and phase of the oscillation correspond to the radius and angle of the dot in polar coordinates. Colours code the oscillators' natural frequencies, running from slowest (red) to fastest (violet). In the absence of coupling, each oscillator would settle onto its limit cycle (circle) and rotate at its natural frequency. However, here all the oscillators are also pulled towards the mean field that they generate collectively (shown as an asterisk at the centre of the population). Time increases from left to right, and from top to bottom. Starting from a random initial condition, the oscillators self-organize by collapsing their amplitudes; then they sort their phases so that the fastest oscillators are in the lead. Ultimately they all rotate as a synchronized pack, with locked amplitudes and phases. The governing equations describe a mean-field model of a laser array²³. (Simulation provided by R. Oliva.)

produce N^2 times as much power as a single one. But in practice, semiconductor laser arrays are notoriously prone to both spatial and temporal instabilities^{20,21}. Even for a simple ring geometry, this problem is dynamically complex.

The first part of this article reviews what is known about dynamical complexity in regular networks of nonlinear systems. I offer a few rules of thumb about the impact of network structure on collective dynamics, especially for arrays of coupled limit-cycle oscillators.

The logical next step would be to tackle networks that combine dynamical and structural complexity, such as power grids or ecological webs. Unfortunately they lie beyond our mathematical reach — we do not even know how to characterize their wiring diagrams. So we have to begin with network topology.

By a happy coincidence, such architectural questions are being pursued in other branches of science, thanks to the excitement about the Internet, functional genomics, financial networks, and so on. The second part of this article uses graph theory to explore the structure of complex networks, an approach that has recently led to some encouraging progress, especially when combined with the tools of statistical mechanics and computer simulations.

Needless to say, many other topics within network science deserve coverage here. The subject is amazingly rich, and apologies are offered to those readers whose favourite topics are omitted.

Regular networks of coupled dynamical systems

Networks of dynamical systems have been used to model everything from earthquakes to ecosystems, neurons to neutrinos^{25–32}. To impose some order on this list, consider the dynamics that each node would exhibit if it were isolated. Assuming it is a generic dynamical

system, its long-term behaviour is given by stable fixed points, limit cycles or chaotic attractors (Box 1).

If we now couple many such systems together, what can be said about their collective behaviour? The answer is not much — the details matter. But I will propose some rough generalizations anyway.

If the dynamical system at each node has stable fixed points and no other attractors, the network tends to lock into a static pattern. Many such patterns may coexist, especially if the nodes have competing interactions. In that case the network may become frustrated and display enormous numbers of locally stable equilibria. This kind of complex static behaviour is seen in models of spin glasses, associative memory neural networks and combinatorial optimization problems³³.

At the opposite extreme, suppose each node has a chaotic attractor. Few rules have emerged about the effect of coupling architecture on dynamics in this case. It is known that networks of identical chaotic systems can synchronize their erratic fluctuations, a curious phenomenon with possible applications to private communications^{34,35}. For a wide range of network topologies, synchronized chaos requires that the coupling be neither too weak nor too strong; otherwise spatial instabilities are triggered³⁴. Related lines of research deal with networks of identical chaotic maps, lattice dynamical systems and cellular automata. These systems have been used mainly as testbeds for exploring spatiotemporal chaos and pattern formation in the simplest mathematical settings, rather than as models of real physical systems.

Identical oscillators

The intermediate case where each node has a stable limit cycle has turned out to be particularly fruitful. Much of the research has been inspired by biological examples, ranging from the mutual synchronization of cardiac pacemaker cells, to rhythmically flashing fireflies and chorusing crickets, to wave propagation in the heart, brain, intestine and nervous system²⁵.

Arrays of identical oscillators often synchronize, or else form patterns that depend on the symmetry of the underlying network³⁶. Other common modes of organization are travelling waves in one spatial dimension, rotating spirals in two dimensions and scroll waves in three dimensions^{25,26}. For fully connected networks where each node is coupled equally to all the others, the completely synchronized state becomes likely.

These heuristics apply to systems coupled by smooth interactions akin to diffusion. But many biological oscillators communicate by sudden impulses: a neuron fires, a firefly flashes, a cricket chirps. Hence the recent interest in pulse-coupled oscillators³⁷. This thread began with Peskin's model of the sinoatrial node, the heart's natural pacemaker, as a collection of N identical integrate-and-fire oscillators³⁸. For the simple case where each oscillator is connected to all the others, Peskin conjectured that they would all end up firing in unison, no matter how they started. He gave a proof for $N=2$ oscillators; it was later demonstrated³⁹ that the conjecture holds for all N . Peskin also conjectured that synchronization would occur even if the oscillators were not quite identical, but that problem remains unproven.

Peskin's model has been used as a caricature of coupled neurons^{40–42} by including synaptic delays, refractory periods, inhibition and local coupling; these realistic features also remove some of the undesirable discontinuities in the mathematics. In an example of scientific cross-fertilization, Hopfield⁴³ pointed out that the locally coupled version of the model is closely related to slider-block models of earthquakes and should therefore display self-organized criticality. That observation suggested intriguing links among neurobiology, geophysics, synchronization and self-organized criticality, and sparked a burst of research activity, as reviewed in ref. 37.

Non-identical oscillators

While modelling populations of biological oscillators, Winfree discovered a new kind of cooperative phenomenon, the temporal

analogue of a phase transition⁴⁴. He proposed a mean-field model of nearly identical, weakly coupled limit-cycle oscillators and showed that when the coupling is small compared to the spread of natural frequencies, the system behaves incoherently, with each oscillator running at its natural frequency. As the coupling is increased, the

incoherence persists until a certain threshold is crossed — then a small cluster of oscillators suddenly ‘freezes’ into synchrony. For still greater coupling, all the oscillators become locked in phase and amplitude (Fig. 2).

Kuramoto²⁶ refined this connection between nonlinear dynamics and statistical physics. He proposed an exactly solvable model of collective synchronization, given by

$$\frac{d\theta_i}{dt} = \omega_i + \frac{K}{N} \sum_{j=1}^N \sin(\theta_j - \theta_i), \quad i = 1, \dots, N$$

where $\theta_i(t)$ is the phase of the i th oscillator and ω_i is its natural frequency, chosen at random from a lorentzian probability density

$$g(\omega) = \frac{\gamma}{\pi[\gamma^2 + (\omega - \omega_0)^2]}$$

of width γ and mean ω_0 . Using an ingenious self-consistency argument, Kuramoto solved for the order parameter

$$r(t) = \left| \frac{1}{N} \sum_{j=1}^N e^{i\theta_j(t)} \right|$$

(a convenient measure of the extent of synchronization) in the limit $N \rightarrow \infty$ and $t \rightarrow \infty$. He found that

$$r = \begin{cases} 0, & K < K_c \\ \sqrt{1 - (K_c/K)}, & K \geq K_c \end{cases}$$

where $K_c = 2\gamma$. In other words, the oscillators are desynchronized completely until the coupling strength K exceeds a critical value K_c . After that, the population splits into a partially synchronized state

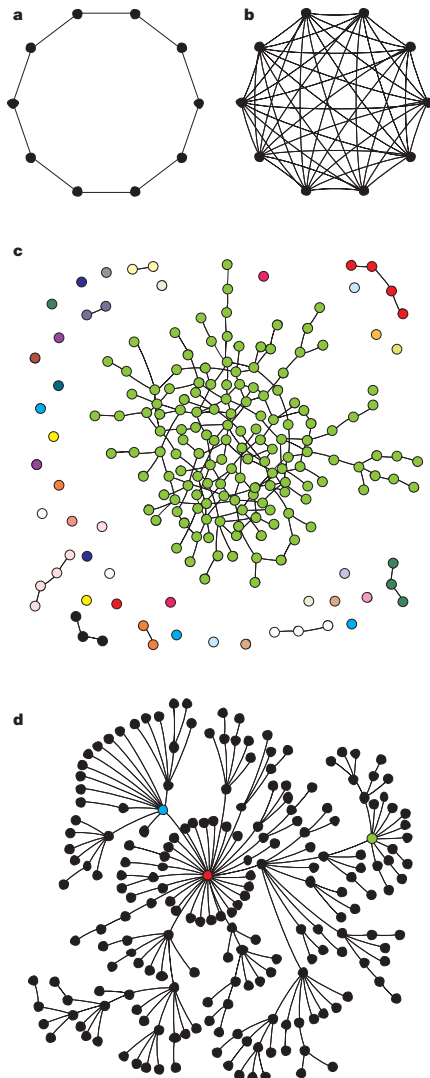


Figure 3 Schematic illustration of regular and random network architectures. **a**, Ring of ten nodes connected to their nearest neighbours. **b**, Fully connected network of ten nodes. **c**, Random graph constructed by placing n nodes on a plane, then joining pairs of them together at random until m links are used. Nodes may be chosen more than once, or not at all. The resulting wiring diagram (not shown) would be a snarl of criss-crossed lines; to clarify it, I have segregated the different connected components, coloured them, and eliminated as many spurious crossings as possible. The main topological features are the presence of a single giant component, as expected^{51–53} for a random graph with $m > n/2$ (here $n = 200$, $m = 193$), and the absence of any dominant hubs. The degree, or number of neighbours, is Poisson distributed across the nodes; most nodes have between one and four neighbours, and all have between zero and six. **d**, Scale-free graph, grown by attaching new nodes at random to previously existing nodes. The probability of attachment is proportional to the degree of the target node; thus richly connected nodes tend to get richer, leading to the formation of hubs and a skewed degree distribution with a heavy tail. Colours indicate the three nodes with the most links (red, $k = 33$ links; blue, $k = 12$; green, $k = 11$). Here $n = 200$ nodes, $m = 199$ links. Figure provided by D. Callaway. Network visualization was done using the Pajek program for large network analysis (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/pajekman.htm>).

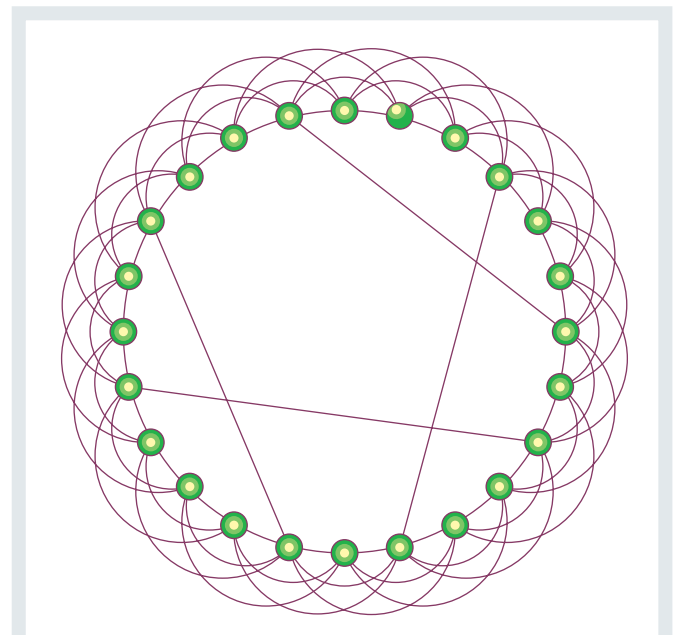


Figure 4 Solvable model of a small-world network. The model starts with a ring lattice of n nodes, each connected to its neighbours out to some range k (here $n = 24$ and $k = 3$). Shortcut links are added between random pairs of nodes, with probability ϕ per link on the underlying lattice. In the limit $n \gg 1$, the average path length between nodes can be approximated analytically. (Adapted from ref. 75.)

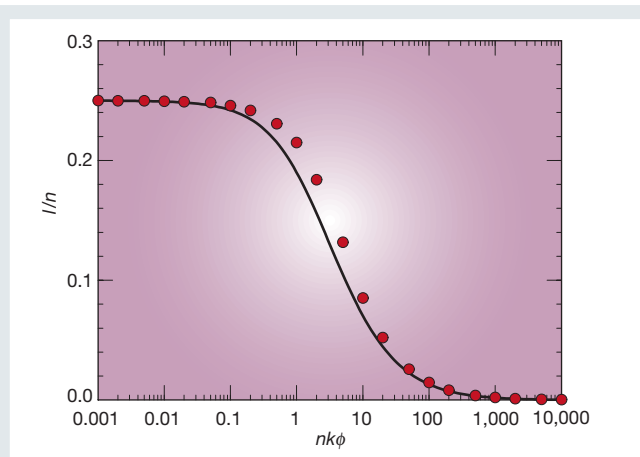


Figure 5 Average path length, normalized by system size, is plotted as a function of the average number of shortcuts. The circles are results from simulations of the model with range $k = 1$ and size up to $n = 10^7$ nodes. The solid line is given by the formula in the text. (Adapted from ref. 75.)

consisting of two groups of oscillators: a synchronized group that contributes to the order parameter r , and a desynchronized group whose natural frequencies lie in the tails of the distribution $g(\omega)$ and are too extreme to be entrained. With further increases in K , more and more oscillators are recruited into the synchronized group, and r grows accordingly.

Twenty-five years later, the Kuramoto model continues to surprise us (see ref. 45 for a review). First, the incoherent state with $r = 0$ was found to be neutrally stable below threshold, despite its apparent stability in simulations; the analysis reveals a connection to Landau damping in plasmas. Second, the square-root critical behaviour of r , almost a cliché for mean-field models in statistical mechanics, turns out to be non-generic; if the sinusoidal coupling is replaced by a periodic function with second harmonics, the scaling changes to $r \sim K - K_c$. Third, although the model was motivated originally by biological oscillators, it has appeared in such far-flung settings as the flavour evolution of neutrinos³², and arrays of Josephson junctions²⁷ and semiconductor lasers²⁴.

The main unsolved problem is the stability of the partially synchronized state for $K > K_c$. Numerical simulations indicate that it is globally stable, in the sense that it attracts almost all solutions, but even the linear stability problem has yet to be solved. Another issue concerns the extension of the model to nearest-neighbour coupling on a d -dimensional cubic lattice. Simulations⁴⁶ and renormalization arguments⁴⁷ indicate that the synchronization phase transition persists for $d \geq 3$ and vanishes for $d = 1$; the results are ambiguous for $d = 2$. All of this awaits a mathematical resolution.

In contrast to the mean-field models of Winfree and Kuramoto, Ermentrout and Kopell's classic work deals with one-dimensional chains of oscillators, first in connection with neuromuscular rhythms in the mammalian intestine⁴⁸, and later in their model of the central pattern generator for the lamprey eel^{49,50}. The main phenomena here involve travelling waves, rather than the synchrony found in mean-field models. This is not accidental, as wave propagation is essential for the generation of peristalsis in the intestine, and for the creation of the swimming rhythm in lamprey.

Ermentrout and Kopell introduced several deep mathematical innovations, but perhaps their most impressive result is a counterintuitive biological prediction. Their lamprey model suggested that the tail-to-head neural connections along the spinal cord would be stronger than those running from head to tail, despite the fact that the wave associated with swimming travels from head to tail. To

everyone's delight, that prediction was later confirmed by their experimental collaborators⁵⁰.

Complex network architectures

All the network topologies discussed so far — chains, grids, lattices and fully-connected graphs — have been completely regular (Fig. 3a, b). Those simple architectures allowed us to focus on the complexity caused by the nonlinear dynamics of the nodes, without being burdened by any additional complexity in the network structure itself. Now I take the complementary approach, setting dynamics aside and turning to more complex architectures. A natural place to start is at the opposite end of the spectrum from regular networks, with graphs that are completely random.

Random graphs

Imagine $n \gg 1$ buttons strewn across the floor⁵¹. Pick two buttons at random and tie them together with thread. Repeat this process m times, always choosing pairs of buttons at random. (If m is large, you might eventually select buttons that already have threads attached. That is certainly allowed; it merely creates clusters of connected buttons.) The result is a physical example of a random graph with n nodes and m links (Fig. 3c). Now slowly lift a random button off the floor. If it is tied to other buttons, either directly or indirectly, those are dragged up too. So what happens? Are you likely to pull up an isolated button, a small cluster or a vast meshwork?

Erdős and Rényi⁵² studied how the expected topology of this random graph changes as a function of m . When m is small, the graph is likely to be fragmented into many small clusters of nodes, called components. As m increases, the components grow, at first by linking to isolated nodes and later by coalescing with other components. A phase transition occurs at $m = n/2$, where many clusters crosslink spontaneously to form a single giant component. For $m > n/2$, this giant component contains on the order of n nodes (more precisely, its size scales linearly with n , as $n \rightarrow \infty$), while its closest rival contains only about $\log n$ nodes. Furthermore, all nodes in the giant component are connected to each other by short paths: the maximum number of 'degrees of separation' between any two nodes grows slowly, like $\log n$.

In the decades since this pioneering work, random graphs have been studied deeply within pure mathematics⁵³. They have also served as idealized coupling architectures for dynamical models of gene networks, ecosystems and the spread of infectious diseases and computer viruses^{29,51,54,55}.

Small-world networks

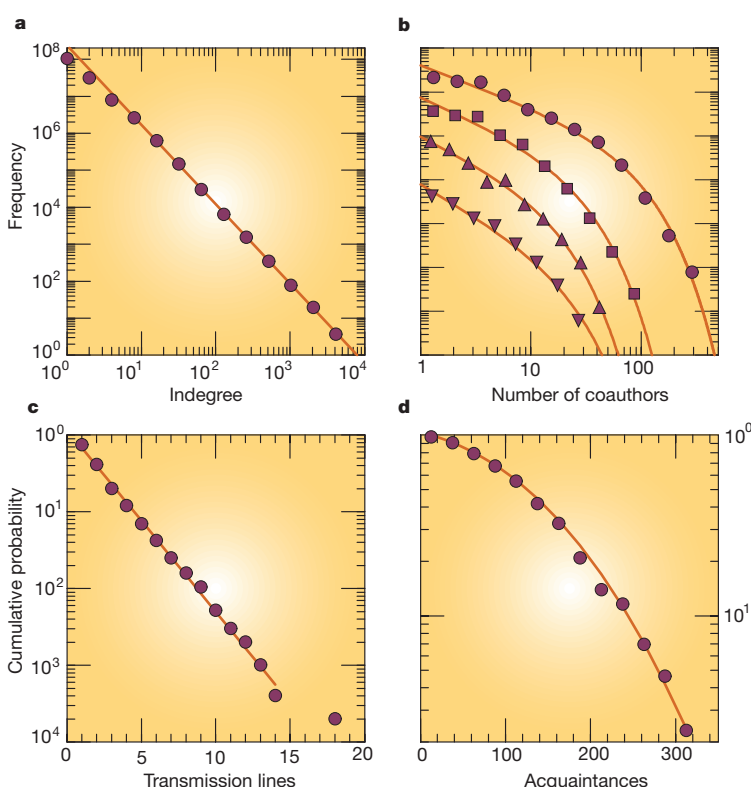
Although regular networks and random graphs are both useful idealizations, many real networks lie somewhere between the extremes of order and randomness. Watts and Strogatz^{56,57} studied a simple model that can be tuned through this middle ground: a regular lattice

Table 1 Clustering for three affiliation networks

Network	Clustering C	
	Theory	Actual
Company directors	0.590	0.588
Movie actors	0.084	0.199
Biomedical authors	0.042	0.088

US corporate directors: 7,673 company directors linked by joint membership on 914 boards of the Fortune 1,000 companies for 1999. Movie actors: 449,913 actors linked by mutual appearances in 151,261 feature films, as specified by the Internet Movie Database (www.imdb.com) as of 1 May 2000. Biomedical collaborations: 1,388,989 scientists linked by coauthorship of at least one of 2,156,769 biomedical journal articles published between 1995 and 1999 inclusive, as listed in the MEDLINE database. The clustering coefficient C is defined as the probability that a connected triple of nodes is actually a triangle; here nodes correspond to people, as in the unipartite representation shown in Fig. 7b. Intuitively, C measures the likelihood that two people who have a mutual collaborator are also collaborators of each other. The results show that the random model accurately predicts C for the corporate director network, given the network's bipartite structure and its degree distributions; no additional social forces need to be invoked. For the networks of actors and scientists, the model accounts for about half of the observed clustering. The remaining portion depends on social mechanisms at work in these communities (see text). (Adapted from ref. 91.)

Figure 6 Degree distributions for real networks. **a**, World-Wide Web. Nodes are web pages; links are directed URL hyperlinks from one page to another. The log–log plot shows the number of web pages that have a given in-degree (number of incoming links). The raw data have been logarithmically binned, with bin ratio 2. The tail of the distribution follows an approximate power law with exponent $\gamma \approx 2.2$. (Adapted from ref. 10.) **b**, Coauthorship networks. Nodes represent scientists; an undirected link between two people means that they have written a paper together. The data are for the years 1995–1999 inclusive and come from three databases: arxiv.org (preprints mainly in theoretical physics), SPIRES (preprints in high-energy physics) and NCSTRL (preprints in computer science). Symbols denote different scientific communities: astrophysics (circles), condensed-matter physics (squares), high-energy physics (upright triangles) and computer science (inverted triangles). The log–log plot shows that the probability of having a total of z coauthors is well approximated by a truncated power law (solid line) of the form $p_z \sim z^{-\gamma} \exp(-z/z_c)$, where z_c is the cutoff. The curves have been displaced vertically for clarity. (Adapted from ref. 14.) **c**, Power grid of the western United States and Canada. Nodes represent generators, transformers and substations; undirected links represent high-voltage transmission lines between them. The data are plotted as a cumulative distribution to reduce the effects of noise on this smaller data set. The linear–log plot shows that the proportion of nodes with at least k transmission lines decays exponentially in k . The negative derivative of this cumulative distribution is the degree distribution, also an exponential. (Adapted from ref. 62.) **d**, Social network. Nodes are 43 Mormons in Utah; undirected links represent acquaintances with other Mormons⁷⁹. The linear–log plot of the cumulative distribution is well fit by an error function (solid line), so the degree distribution is a gaussian. (Adapted from ref. 62.)



where the original links are replaced by random ones with some probability $0 \leq \phi \leq 1$. They found that the slightest bit of rewiring transforms the network into a ‘small world’, with short paths between any two nodes, just as in the giant component of a random graph. Yet the network is much more highly clustered than a random graph, in the sense that if A is linked to B and B is linked to C , there is a greatly increased probability that A will also be linked to C (a property that sociologists⁵⁸ call ‘transitivity’).

Watts and Strogatz conjectured that the same two properties — short paths and high clustering — would hold also for many natural and technological networks. Furthermore, they conjectured that dynamical systems coupled in this way would display enhanced signal propagation speed, synchronizability and computational power, as compared with regular lattices of the same size. The intuition is that the short paths could provide high-speed communication channels between distant parts of the system, thereby facilitating any dynamical process (like synchronization or computation) that requires global coordination and information flow.

Research has proceeded along several fronts. Many empirical examples of small-world networks have been documented, in fields ranging from cell biology to business^{9,14,59–64}. On the theoretical side, small-world networks are turning out to be a Rorschach test — different scientists see different problems here, depending on their disciplines.

Computer scientists see questions about algorithms and their complexity. Walsh⁶⁵ showed that graphs associated with many difficult search problems have a small-world topology. Kleinberg⁶⁶ introduced an elegant model of the algorithmic challenge posed by Milgram’s original sociological experiment⁶⁷ — how to actually find a short chain of acquaintances linking yourself to a random target person, using only local information — and he proved that the problem is easily solvable for some kinds of small worlds, and essentially intractable for others.

Epidemiologists have asked how local clustering and global

contacts together influence the spread of infectious disease, with implications for vaccination strategies and the evolution of virulence^{68–71}. Neurobiologists have wondered about the possible evolutionary significance of small-world neural architecture. They have argued that this kind of topology combines fast signal processing with coherent oscillations⁷², unlike either regular or random architectures, and that it may be selected by adaptation to rich sensory environments and motor demands⁶⁴.

Perhaps the strongest response to the Rorschach test comes from the statistical physicists, who sensed immediately⁷³ that the toy model of Watts and Strogatz⁵⁶ would yield to their techniques (see ref. 74 for a review). In its improved form the model starts with a ring of n nodes, each connected by undirected links to its nearest and next-nearest neighbours out to some range k . Shortcut links are then added — rather than rewired — between randomly selected pairs of nodes, with probability ϕ per link on the underlying lattice; thus there are typically $nk\phi$ shortcuts in the graph (Fig. 4). The question is: on average, how many steps are required to go from one node to another along the shortest route? If ℓ denotes that average separation, we find that ℓ drops sharply near $\phi=0$, confirming that a few shortcuts do indeed shrink the world dramatically. One of the most striking results is the following formula derived by Newman, Moore and Watts⁷⁵:

$$\ell = \frac{n}{k} f(nk\phi)$$

where

$$f(x) = \frac{1}{2\sqrt{x^2 + 2x}} \tanh^{-1} \frac{x}{\sqrt{x^2 + 2x}}$$

This solution is asymptotically exact in the limits $n \rightarrow \infty$ (large system size) and either $nk\phi \rightarrow \infty$ or $nk\phi \rightarrow 0$ (large or small number of

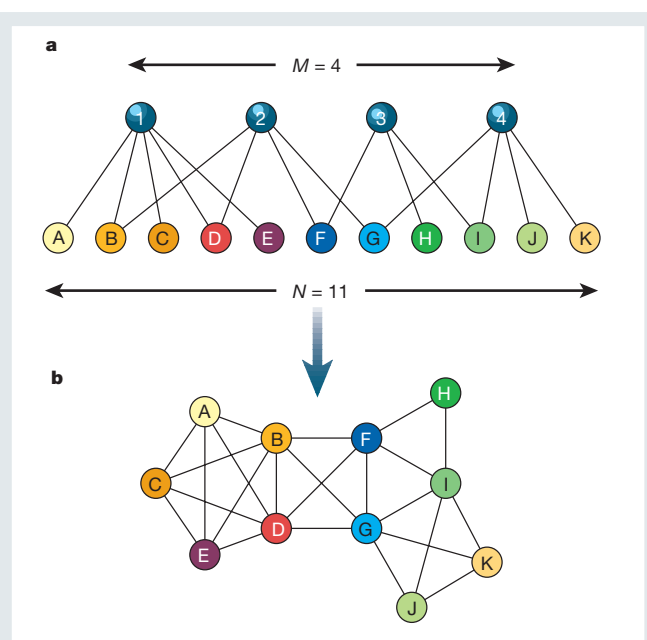


Figure 7 Bipartite versus unipartite representations of the corporate director network. **a**, In the bipartite approach, directors and boards are treated as distinct kinds of nodes. The schematic example shows 11 directors and 4 boards. Links indicate which people sit on which boards. By definition there are no links between pairs of people, or between pairs of boards. **b**, The more familiar unipartite representation depicts the people as nodes, with links between those on the same board, forming cliques. This approach discards important information and can conflate different structures. For example, the triangle FHI corresponds to board number 3, as seen in **a**, whereas the similar-looking triangle FGI does not correspond to any single board. Another confusing effect is the large number of cliques that occur automatically in this projection of the full bipartite graph. Such cliques account for much of the high clustering observed in real affiliation networks⁵⁹. The random graph model teases out this generic source of clustering from that indicative of more interesting social interactions. (Adapted from ref. 91.)

shortcuts). Figure 5 shows that it also gives the correct qualitative behaviour for $nk\phi \approx 1$. Barbour and Reinert⁷⁶ improved this result by proving a rigorous distributional approximation for ℓ , along with a bound on the error.

Scale-free networks

In any real network, some nodes are more highly connected than others are. To quantify this effect, let p_k denote the fraction of nodes that have k links. Here k is called the degree and p_k is the degree distribution.

The simplest random graph models^{52,53} predict a bell-shaped Poisson distribution for p_k . But for many real networks, p_k is highly skewed and decays much more slowly than a Poisson. For instance, the distribution decays as a power law $p_k \sim k^{-\gamma}$ for the Internet backbone¹¹, metabolic reaction networks⁹, the telephone call graph¹³ and the World-Wide Web¹⁰ (Fig. 6a). Remarkably, the exponent $\gamma \approx 2.1$ – 2.4 for all of these cases. Taken literally, this form of heavy-tailed distribution would imply an infinite variance. In reality, there are a few nodes with many links (Fig. 3d). For the World-Wide Web, think Yahoo; for metabolic networks, think ATP. Barabási, Albert and Jeong^{77,78} have dubbed these networks ‘scale-free’, by analogy with fractals, phase transitions and other situations where power laws arise and no single characteristic scale can be defined.

The scale-free property is common but not universal⁶². For coauthorship networks of scientists, p_k is fit better by a power law with an exponential cutoff⁴ (Fig. 6b); for the power grid of the western United States, p_k is an exponential distribution⁶² (Fig. 6c); and for a social network of Mormons in Utah⁷⁹, p_k is gaussian⁶² (Fig. 6d).

Nevertheless, the scale-free case has stimulated a great deal of theorizing. The earliest work is due to Simon^{80,81} in 1955, now independently rediscovered by Barabási, Albert and Jeong^{77,78}. They showed that a heavy-tailed degree distribution emerges automatically from a stochastic growth model in which new nodes are added continuously and attach themselves preferentially to existing nodes, with probability proportional to the degree of the target node. Richly connected nodes get richer, and the result is $p_k \sim k^{-3}$. More sophisticated models^{82–84} include the effects of adding or rewiring links, allowing nodes to age so that they can no longer accept new links, or varying the form of preferential attachment. These generalized models predict exponential and truncated power-law p_k in some parameter regimes, as well as scale-free distributions.

Could there be a functional advantage to scale-free architecture? Albert, Jeong and Barabási⁸⁵ suggested that scale-free networks are resistant to random failures because a few hubs dominate their topology (Fig. 3d). Any node that fails probably has small degree (like most nodes) and so is expendable. The flip side is that such networks are vulnerable to deliberate attacks on the hubs. These intuitive ideas have been confirmed numerically^{10,85} and analytically^{86,87} by examining how the average path length and size of the giant component depend on the number and degree of the nodes removed. Some possible implications for the resilience of the Internet^{79–81}, the design of therapeutic drugs⁹, and the evolution of metabolic networks^{9,59} have been discussed.

Generalized random graphs

As mentioned above, the simplest random graph predicts a Poisson degree distribution, and so cannot accommodate the other types of distribution found in real networks. Molloy and Reed^{88,89} introduced a more flexible class of random graphs in which any degree distribution is permitted. Given a sequence of non-negative integers $\{d_k\}$, where d_k denotes the number of nodes with degree k , consider the ensemble of all graphs with that prescribed degree sequence, and weight them all equally when computing statistical averages of interest. For this class of graphs, Molloy and Reed derived a simple condition for the birth of the giant component⁸⁸, and they also found an implicit formula for its size as a fraction of n , the total number of nodes⁸⁹. Specifically, let $n \gg 1$ and define

$$Q = \sum_{k=1}^{\infty} p_k k(k-2)$$

where $p_k = d_k/n$. If $Q < 0$, the graph consists of many small components. The average component size diverges as $Q \rightarrow 0$ from below, and a giant component exists for $Q > 0$. (In technical terms, these results hold ‘almost surely’; that is, with probability tending to 1 as $n \rightarrow \infty$.)

Aiello, Chung and Lu⁹⁰ applied these results to a random graph model for scale-free networks. For p_k of power-law form, the condition on Q implies that a giant component exists if and only if $\gamma < 3.47$, which holds for most scale-free networks measured so far. If $\gamma < 1$, there are so many high-degree hubs that the network forms one huge, connected piece. They also proved theorems about the number and size of small components outside the giant component, and compared these to a real graph of about 47 million telephone numbers and the calls between them in one day. They found that the data are best fit by an exponent $\gamma \approx 2.1$, which predicts correctly that the call graph is not connected but has a giant component.

The papers by Molloy and Reed^{88,89} and Aiello *et al.*⁹⁰ are mathematically rigorous. Newman, Strogatz and Watts⁹¹ recently developed a more heuristic approach based on generating functions. By handling the limit $n \rightarrow \infty$ in an intuitive way, their approach yields elementary derivations of the earlier results, along with new exact results for graphs with additional structure, such as directed or bipartite graphs.

The bipartite case is especially interesting for applications⁵⁸. By definition, in a bipartite graph there are two types of nodes, with links running only between different kinds (Fig. 7). For example, consider the network of the boards of directors of the Fortune 1,000 companies, the largest US corporations ranked according to revenues. This network is fascinating because the boards ‘interlock’ — some important people sit on several of them — and this overlap knits virtually all large US firms together into a giant web of corporate governance¹⁷.

Let p_j denote the probability that a director sits on exactly j boards, and let q_k denote the probability that a board consists of k directors. Figure 8 shows that p_j is approximately exponential, with most directors sitting on only one board, whereas q_k is strongly peaked around $k = 10$, indicating that most boards have about ten members. As a null hypothesis, assume that the Fortune 1,000 network is a random member of the ensemble of all bipartite graphs with the same p_j and q_k . Then generating functions yield predictions for various quantities of interest⁹¹. For example, suppose we want to calculate r_z , the probability that a random director works with a total of z other co-directors, summed over all the boards on which he or she serves. Let

$$f_0(x) = \sum_{j=0}^{\infty} p_j x^j$$

$$g_0(x) = \sum_{k=0}^{\infty} q_k x^k$$

be the generating functions associated with the empirical degree distributions p_j and q_k . If we now choose a random edge on the bipartite graph and follow it to the board at one of its ends, the distribution of the number of other edges leaving that board can be shown to be generated by $g_1(x) = g_0'(x)/\nu$, where $\nu = g_0'(1)$. Then for a randomly chosen director, the generating function for z is given by $G_0(x) = f_0(g_1(x))$. If we expand G_0 in a series as

$$G_0(x) = \sum_{z=0}^{\infty} r_z x^z,$$

the coefficients r_z are exactly the quantities we seek. They can be extracted by repeated differentiation: $r_z = (1/z!)(d^z G_0/dx^z)|_{x=0}$.

Figure 8c shows that the predicted r_z agrees almost perfectly with the actual distribution. Similarly, the clustering coefficient⁵⁶ predicted for the directors lies within 1% of the observed value (Table 1). Clearly the random model captures much of the structure of the real network.

However, for two other bipartite graphs — film actors and the movies they appeared in, and biomedical scientists and the papers they coauthored — the model⁹¹ underestimates the clustering coefficients by half (Table 1). The reason is that the random model quantifies only the generic portion of the clustering; it reflects the cliques that are formed automatically whenever a bipartite collaboration graph is projected onto the space of people, as in Fig. 7b. For the corporate board data, those cliques account for essentially all the clustering (simply because most directors sit on only one board, thus preventing clustering across boards). But for the scientists and actors, some further mechanisms must be at work. One possible explanation is that scientists tend to introduce pairs of their collaborators to each other, engendering new collaborations.

In this way the random model allows us to disentangle the generic features of bipartite graphs from those that could reflect sociological effects. Beyond their benchmarking role, generalized random graphs provide a promising new class of substrates on which dynamical processes can be simulated and even approached analytically. Using this approach, Watts⁹² has given an intriguing explanation of fads and

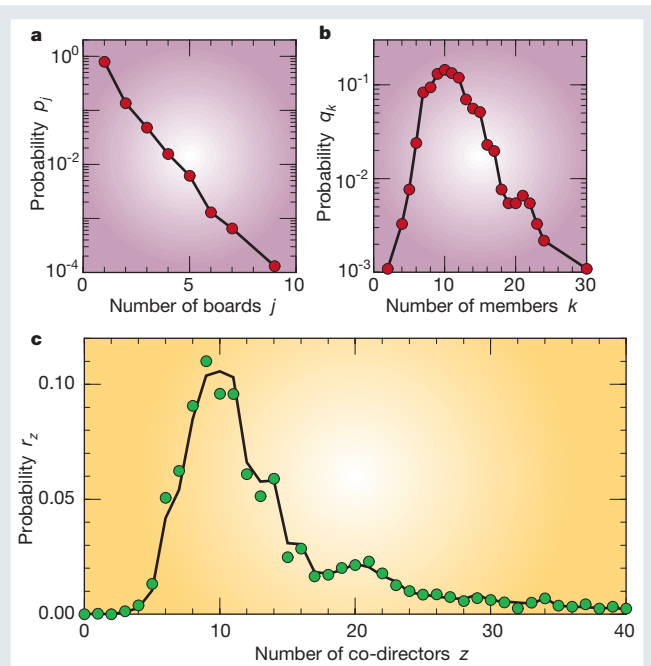


Figure 8 Structural properties of the Fortune 1,000 network of corporate directors for 1999. The data shown here comprise 7,673 directors and 914 boards. Most directors serve on the board of only one company, but about 20% sit on two or more boards, thereby creating interlocked directorates. This has potentially important consequences for business practices in the United States¹⁷. **a**, Distribution of the number of boards per director. The probability p_j that a director sits on exactly j boards decreases roughly exponentially with j . **b**, Distribution of the number of directors per board. The probability q_k that a board has k members is approximately lognormally distributed, with a typical board size around $k = 10$ members. **c**, Distribution of each director's total number of co-directors, summed over all the boards on which the director sits. The probability r_z of serving with z other co-directors is a complicated function of z , with a long tail and a secondary shoulder near $z = 20$, yet the theoretical prediction (solid line) agrees almost perfectly with the actual distribution (green circles). (Adapted from ref. 91.)

normal accidents as the natural consequence of cascade dynamics on sparse interaction networks.

Outlook

In the short run there are plenty of good problems about the nonlinear dynamics of systems coupled according to small-world, scale-free or generalized random connectivity. The speculations that these architectures are dynamically advantageous (for example, more synchronizable or error-tolerant) need to be sharpened, then confirmed or refuted mathematically for specific examples. Other ripe topics include the design of self-healing networks, and the relationships among optimization principles, network growth rules and network topology^{82–84,93–96}.

In the longer run, network thinking will become essential to all branches of science as we struggle to interpret the data pouring in from neurobiology, genomics, ecology, finance and the World-Wide Web. Will theory be able to keep up? Time to log back on to the Internet... □

1. Western Systems Coordinating Council (WSCC). Disturbance Report for the Power System Outage that Occurred on the Western Interconnection on August 10th, 1996 at 1548 PAST <<http://www.wscoc.com>> (October 1996).
2. Anonymous. Media: Six degrees from Hollywood. *Newsweek* 11 October 1999, 6 (1999).
3. Kirby, D. & Sahre, P. Six degrees of Monica. *New York Times* 21 February 1998, op. ed. page (1998).
4. Cohen, J. E., Briand, F. & Newman, C. M. *Community Food Webs: Data and Theory* (Springer, Berlin, 1990).
5. Williams, R. J. & Martinez, N. D. Simple rules yield complex food webs. *Nature* **404**, 180–183 (2000).
6. Kohn, K. W. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Mol. Biol. Cell* **10**, 2703–2734 (1999).

7. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47–C52 (1999).
8. Bhalla, U. S. & Iyengar, R. Emergent properties of networks of biological signalling pathways. *Science* **283**, 381–387 (1999).
9. Jeong H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
10. Broder, A. *et al.* Graph structure in the web. *Comput. Netw.* **33**, 309–320 (2000).
11. Faloutsos, M., Faloutsos, P. & Faloutsos, C. On power-law relationships of the internet topology. *Comp. Comm. Rev.* **29**, 251–262 (1999).
12. Achacoso, T. B. & Yamamoto, W. S. *AY's Neuroanatomy of C. elegans for Computation* (CRC Press, Boca Raton, FL, 1992).
13. Abello, J., Buchsbaum, A. & Westbrook, J. A functional approach to external graph algorithms. *Lect. Notes Comput. Sci.* **1461**, 332–343 (1998).
14. Newman, M. E. J. The structure of scientific collaboration networks. *Proc. Natl Acad. Sci. USA* **98**, 404–409 (2001).
15. Seglen, P. O. The skewness of science. *J. Am. Soc. Inform. Sci.* **43**, 628–638 (1992).
16. Redner, S. How popular is your paper? An empirical study of the citation distribution. *Eur. J. Phys. B* **4**, 131–134 (1998).
17. Davis, G. F. The significance of board interlocks for corporate governance. *Corp. Govern.* **4**, 154–159 (1996).
18. Wilson, E. O. *Consilience* p.85 (Knopf, New York, 1998).
19. Weiss, C. O. & Vilaseca, R. *Dynamics of Lasers* (VCH, Weinheim, 1991).
20. Winful, H. G. & Wang, S. S. Stability of phase locking in coupled semiconductor laser arrays. *Appl. Phys. Lett.* **53**, 1894–1896 (1988).
21. Li, R. D. & Erneux, P. Preferential instability in arrays of coupled lasers. *Phys. Rev. A* **46**, 4252–4260 (1992).
22. Fabiny, L., Colet, P., Roy, R. & Lenstra, D. Coherence and phase dynamics of spatially coupled solid-state lasers. *Phys. Rev. A* **47**, 4287–4296 (1993).
23. Kourtchatov, S. Yu., Likhanskii, V. V., Naporotovich, A. P., Arecchi, F. T. & Lapucci, A. Theory of phase locking of globally coupled laser arrays. *Phys. Rev. A* **52**, 4089–4094 (1995).
24. Kozzyreff, G., Vladimirov, A. G. & Mandel, P. Global coupling with time delay in an array of semiconductor lasers. *Phys. Rev. Lett.* **85**, 3809–3812 (2000).
25. Winfree, A. T. *The Geometry of Biological Time* (Springer, New York, 1980).
26. Kuramoto, Y. *Chemical Oscillations, Waves, and Turbulence* (Springer, Berlin, 1984).
27. Wiesenfeld, K., Colet, P. & Strogatz, S. H. Frequency locking in Josephson arrays: connection with the Kuramoto model. *Phys. Rev. E* **57**, 1563–1569 (1998).
28. Turcotte, D. L. *Fractals and Chaos in Geology and Geophysics* 2nd edn (Cambridge Univ. Press, Cambridge, 1997).
29. May, R. M. *Stability and Complexity in Model Ecosystems* (Princeton Univ. Press, Princeton, 1973).
30. Levin, S. A., Grenfell, B. T., Hastings, A. & Perelson, A. S. Mathematical and computational challenges in population biology and ecosystem science. *Science* **275**, 334–343 (1997).
31. Arbib, M. (ed.) *The Handbook of Brain Theory and Neural Networks* (MIT Press, Cambridge, MA, 1995).
32. Pantaleone, J. Stability of incoherence in an isotropic gas of oscillating neutrinos. *Phys. Rev. D* **58**, 3002 (1998).
33. Stein, D. L. (ed.) *Lectures in the Sciences of Complexity* (Addison-Wesley, Reading, MA, 1989).
34. Pecora, L. M., Carroll, T. L., Johnson, G. A., Mar, D. J. & Heagy, J. F. Fundamentals of synchronization in chaotic systems: concepts and applications. *Chaos* **7**, 520–543 (1997).
35. VanWiggeren, G. D. & Roy, R. Communication with chaotic lasers. *Science* **279**, 1198–1200 (1998).
36. Collins, J. J. & Stewart, I. Coupled nonlinear oscillators and the symmetries of animal gaits. *J. Nonlin. Sci.* **3**, 349–392 (1993).
37. Pérez, C. J., Corral, A., Díaz-Guilera, A., Christensen, K. & Arenas, A. On self-organized criticality and synchronization in lattice models of coupled dynamical systems. *Int. J. Mod. Phys. B* **10**, 1111–1151 (1996).
38. Peskin, C. S. *Mathematical Aspects of Heart Physiology* 268–278 (Courant Institute of Mathematical Sciences, New York, 1975).
39. Mirollo, R. E. & Strogatz, S. H. Synchronization of pulse-coupled biological oscillators. *SIAM J. Appl. Math.* **50**, 1645–1662 (1990).
40. Abbott, L. F. & van Vreeswijk, C. Asynchronous states in neural networks of pulse-coupled oscillators. *Phys. Rev. E* **48**, 1483–1490 (1993).
41. Bressloff, P. C. Mean-field theory of globally coupled integrate-and-fire neural oscillators with dynamic synapses. *Phys. Rev. E* **60**, 2160–2170 (1999).
42. Golomb, D. & Hansel, D. The number of synaptic inputs and the synchrony of large, sparse neuronal networks. *Neural Comput.* **12**, 1095–1139 (2000).
43. Hopfield, J. J. Neurons, dynamics, and computation. *Phys. Today* **47**, 40–46 (1994).
44. Winfree, A. T. Biological rhythms and the behavior of populations of coupled oscillators. *J. Theor. Biol.* **16**, 15–42 (1967).
45. Strogatz, S. H. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D* **143**, 1–20 (2000).
46. Sakaguchi, H., Shinomoto, S. & Kuramoto, Y. Local and global self-entrainments in oscillator lattices. *Prog. Theor. Phys.* **77**, 1005–1010 (1987).
47. Daido, H. Lower critical dimension for populations of oscillators with randomly distributed frequencies: a renormalization-group analysis. *Phys. Rev. Lett.* **61**, 231–234 (1988).
48. Ermentrout, G. B. & Kopell, N. Frequency plateaus in a chain of weakly coupled oscillators. *SIAM J. Math. Anal.* **15**, 215–237 (1984).
49. Kopell, N. & Ermentrout, G. B. Symmetry and phaselocking in chains of weakly coupled oscillators. *Commun. Pure Appl. Math.* **39**, 623–660 (1986).
50. Sigvardt, K. A. & Williams, T. L. Models of central pattern generators as oscillators: the lamprey locomotor CPG. *Semin. Neurosci.* **4**, 37–46 (1992).
51. Kauffman, S. *At Home in the Universe* (Oxford, New York, 1995).
52. Erdős, P. & Rényi, A. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–61 (1960).
53. Bollobás, B. *Random Graphs* (Academic, London, 1985).
54. Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467 (1969).
55. Kephart, J. O. & White, S. R. in *Proc. 1991 IEEE Comput. Soc. Symp. Res. Security Privacy* 343–359 (IEEE Computer Society Press, Los Alamitos, CA, 1991).
56. Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440–442 (1998).
57. Watts, D. J. *Small Worlds* (Princeton Univ. Press, Princeton 1999).
58. Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications* (Cambridge Univ. Press, New York, 1994).
59. Wagner, A. & Fell, D. The small world inside large metabolic networks. Preprint available at <http://www.santafe.edu/sfi/publications/Abstracts/00-07-041abs.html> (2000).
60. Adamic, L. The small world web. *Lect. Notes Comput. Sci.* **1696**, 443–452 (Springer, New York, 1999).
61. Kogut, B. & Walker, G. Small worlds and the durability of national networks: ownership and acquisitions in Germany. *Am. Sociol. Rev.* (in the press).
62. Amaral, L. A. N., Scala, A., Barthélémy, M. & Stanley, H. E. Classes of behavior of small-world networks. *Proc. Natl Acad. Sci. USA* **97**, 11149–11152 (2000).
63. Stephan, K. E. *et al.* Computational analysis of functional connectivity between areas of primate visual cortex. *Phil. Trans. R. Soc. Lond. B* **355**, 111–126 (2000).
64. Sporns, O., Tononi, G. & Edelman, G. M. Theoretical neuroanatomy: relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cereb. Cortex* **10**, 127–141 (2000).
65. Walsh, T. in *Proc. 16th Int. Joint Conf. Artif. Intell.* 1172–1177 <http://dream.dai.ed.ac.uk/group/tw/papers/wijcai99.ps>
66. Kleinberg, J. M. Navigation in a small world. *Nature* **406**, 845 (2000).
67. Milgram, S. The small world problem. *Psychol. Today* **2**, 60–67 (1967).
68. Wallinga, J., Edmunds, K. J. & Kretzschmar, M. Perspective: human contact patterns and the spread of airborne infectious diseases. *Trends Microbiol.* **7**, 372–377 (1999).
69. Ball, F., Mollison, J. & Scalia-Tomba, G. Epidemics with two levels of mixing. *Ann. Appl. Probab.* **7**, 46–89 (1997).
70. Keeling, M. J. The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. Lond. B* **266**, 859–867 (1999).
71. Boots, M. & Sasaki, A. ‘Small worlds’ and the evolution of virulence: infection occurs locally and at a distance. *Proc. R. Soc. Lond. B* **266**, 1933–1938 (1999).
72. Lago-Fernandez, L. F., Huerta, R., Corbacho, F. & Sigüenza, J. Fast response and temporal coherent oscillations in small-world networks. *Phys. Rev. Lett.* **84**, 2758–2761 (2000).
73. Barthélémy, M. & Amaral, L. A. N. Small-world networks: evidence for a crossover picture. *Phys. Rev. Lett.* **82**, 3180–3183 (1999).
74. Newman, M. E. J. Models of the small world: a review. *J. Stat. Phys.* **101**, 819–841 (2000).
75. Newman, M. E. J., Moore, C. & Watts, D. J. Mean-field solution of the small-world network model. *Phys. Rev. Lett.* **84**, 3201–3204 (2000).
76. Barbour, A. D. & Reinert, G. Small worlds. Preprint cond-mat/0006001 at <http://xxx.lanl.gov> (2000).
77. Barabási, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
78. Barabási, A.-L., Albert, R. & Jeong, H. Mean-field theory for scale-free random networks. *Physica A* **272**, 173–197 (1999).
79. Bernard, H. R., Killworth, P. D., Evans, M. J., McCarty, C. & Shelley, G. A. Studying social relations cross-culturally. *Ethnology* **27**, 155–179 (1988).
80. Simon, H. A. On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955).
81. Bornholdt, S. & Ebel, H. World-Wide Web scaling exponent from Simon’s 1955 model. Preprint cond-mat/0008465 at <http://xxx.lanl.gov> (2000).
82. Albert, R. & Barabási, A.-L. Topology of evolving networks: local events and universality. *Phys. Rev. Lett.* **85**, 5234–5237 (2000).
83. Dorogovtsev, S. N. & Mendes, J. F. F. Evolution of networks with aging of sites. *Phys. Rev. E* **62**, 1842–1845 (2000).
84. Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of growing random networks. *Phys. Rev. Lett.* **85**, 4629–4632 (2000).
85. Albert, R., Jeong, H. & Barabási, A.-L. Error and attack tolerance of complex networks. *Nature* **406**, 378–382 (2000).
86. Cohen, R., Erez, K., ben-Avraham, D. & Havlin, S. Resilience of the Internet to random breakdowns. *Phys. Rev. Lett.* **85**, 4626–4628 (2000).
87. Callaway, D. S., Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Network robustness and fragility: percolation on random graphs. *Phys. Rev. Lett.* **85**, 5468–5471 (2000).
88. Molloy, M. & Reed, B. A critical point for random graphs with given degree sequence. *Random Struct. Algorithms* **6**, 161–179 (1995).
89. Molloy, M. & Reed, B. The size of the giant component of a random graph with given degree sequence. *Combinatorics Probab. Comput.* **7**, 295–305 (1998).
90. Aiello, W., Chung, F. & Lu, L. A random graph model for power law graphs. *Exp. Math.* (in the press); preprint available at <http://math.ucsd.edu/~fan/power.pdf>.
91. Newman, M. E. J., Watts, D. J. & Strogatz, S. H. Random graphs with arbitrary degree distribution and their applications. Preprint cond-mat/0007235 at <http://xxx.lanl.gov> (2000).
92. Watts, D. J. A simple model of fads and cascading failures. Preprint available at <http://www.santafe.edu/sfi/publications/Abstracts/00-12-062abs.html> (2000).
93. Cheriak, C. Component placement optimization in the brain. *J. Neurosci.* **14**, 2418–2427 (1994).
94. Mitchison, G. Neuronal branching patterns and the economy of cortical wiring. *Proc. R. Soc. Lond. B* **245**, 151–158 (1991).
95. West, G. B., Brown, J. H. & Enquist, B. J. The fourth dimension of life: fractal geometry and the allometric scaling of organisms. *Science* **284**, 1677–1679 (1999).
96. Banavar, J. R., Colaiori, F., Flammini, A., Maritan, A. & Rinaldo, A. Topology of the fittest transportation network. *Phys. Rev. Lett.* **84**, 4745–4748 (2000).
97. Strogatz, S. H. *Nonlinear Dynamics and Chaos* (Perseus, New York, 1994).

Acknowledgements

Thanks to J. Ariaratnam, A.-L. Barabási, N. Martinez, M. E. J. Newman, D. Watts and A. Winfree for their comments on a draft of the manuscript, and to R. Albert, L. Amaral, M. Amin, W. Blake, A. Broder, D. Callaway, J. Collins, G. Davis, H. Ebel, K. Kohn, N. Martinez, R. Oliva, M. E. J. Newman, J. Thorp, D. Watts, J. Wiener, A. Winfree and H. Wang for providing data, figures and information. Research supported in part by the National Science Foundation, Department of Defense, and Electric Power Research Institute.

Synchronization and rhythmic processes in physiology

Leon Glass

Department of Physiology, Centre for Nonlinear Dynamics in Physiology and Medicine, McGill University, Montreal, Quebec, Canada H3G 1Y6

Complex bodily rhythms are ubiquitous in living organisms. These rhythms arise from stochastic, nonlinear biological mechanisms interacting with a fluctuating environment. Disease often leads to alterations from normal to pathological rhythm. Fundamental questions concerning the dynamics of these rhythmic processes abound. For example, what is the origin of physiological rhythms? How do the rhythms interact with each other and the external environment? Can we decode the fluctuations in physiological rhythms to better diagnose human disease? And can we develop better methods to control pathological rhythms? Mathematical and physical techniques combined with physiological and medical studies are addressing these questions and are transforming our understanding of the rhythms of life.

Physiological rhythms are central to life. We are all familiar with the beating of our hearts, the rhythmic motions of our limbs as we walk, our daily cycle of waking and sleeping, and the monthly menstrual cycle. Other rhythms, equally important but not as obvious, underlie the release of hormones regulating growth and metabolism, the digestion of food, and many other bodily processes. The rhythms interact with each other as well as the outside fluctuating, noisy environment under the control of innumerable feedback systems that provide an orderly function that enables life. Disruption of the rhythmic processes beyond normal bounds or emergence of abnormal rhythms is associated with disease. Figure 1 shows several examples of complex physiological rhythms.

The investigation of the origin and dynamics of these rhythmic processes — once the sole province of physicians and experimental physiologists — is coming under increasingly close examination by mathematicians and physicists. Mathematical analyses of physiological rhythms show that nonlinear equations (see Box 1) are necessary to describe physiological systems^{1–4}. In contrast to the linear equations of traditional mathematical physics (for example, Maxwell's equations, the heat equation, the wave equation or Schrödinger's equation), nonlinear equations rarely admit an analytical solution. Consequently, as Hodgkin and Huxley realized in their classic analysis of the properties of ionic channels in the membranes of squid nerve cells, numerical simulations are one essential feature of quantitative studies of physiological systems⁵. A complementary approach is to analyse qualitative aspects of simplified mathematical models of physiological systems. This involves a mathematical analysis of those features of physiological systems that will be preserved by classes of models that are sufficiently close to the real system. For example, periodic stimulation of a squid giant axon gives rise to a wide variety of regular and irregular rhythms that can be modelled by simple as well as complex mathematical models^{6–8}.

Although we know that deterministic equations can display chaotic dynamics, it is not straightforward to distinguish deterministic 'chaotic' dynamics from 'noisy' dynamics in real experimental data. The problems were underscored by Ruelle: "Real systems can in general be described as deterministic systems with some added

noise"⁹. Although in some carefully controlled situations it is possible to obtain good evidence that a system is obeying deterministic equations with a small amount of noise^{6–8,10}, more usually the origin and the amount of 'noise' is not easy to determine. In this review, I concentrate on three fundamental issues related to synchronization and rhythmic processes in physiology: origins of complex physiological rhythms; synchronization of physiological oscillations; and the function of noise and chaos in physiological processes with particular emphasis on stochastic resonance. Finally, I discuss the potential applications of these ideas to medicine.

Origins of complex physiological rhythms

Physiological rhythms are rarely strictly periodic but rather fluctuate irregularly over time (Fig. 1). The fluctuations arise from the combined influences of the fluctuating environment, the 'noise' that is inherent in biological systems, and deterministic, possibly chaotic, mechanisms. In most natural as opposed to laboratory settings, there is continual interaction between the environment and the internal control mechanisms, so that separation of dynamics due to intrinsic rather than extrinsic mechanism is not possible. Independent of the mechanism for the fluctuation, it is usually not clear whether the fluctuations are essential to the physiological function, or whether the physiological functions are carried out despite the fluctuation.

At a subcellular level, ionic channels in cell membranes open and close in response to voltage and chemical changes in a cell's environment. An ionic channel lets a specific ion pass through it provided there is a concentration gradient of that ion between the intracellular and extracellular medium and the channel is open. Because histograms of open and closed times of ionic channels are often well fit by an exponential or a sum of exponentials, theoretical models of channel activity often assume that the dynamics of channel opening and closing are governed by simple random processes such as the Poisson process^{11–13}. Figure 2a shows a schematic representation of five channels, each of which is open at time 0 and each of which closes randomly with a fixed probability of 0.1 ms^{-1} . Figure 2b shows the fraction of open channels as a function of time for a membrane with 5, 50 and 500 channels. As the numbers of channels in a membrane increases, the falloff of the fraction of open

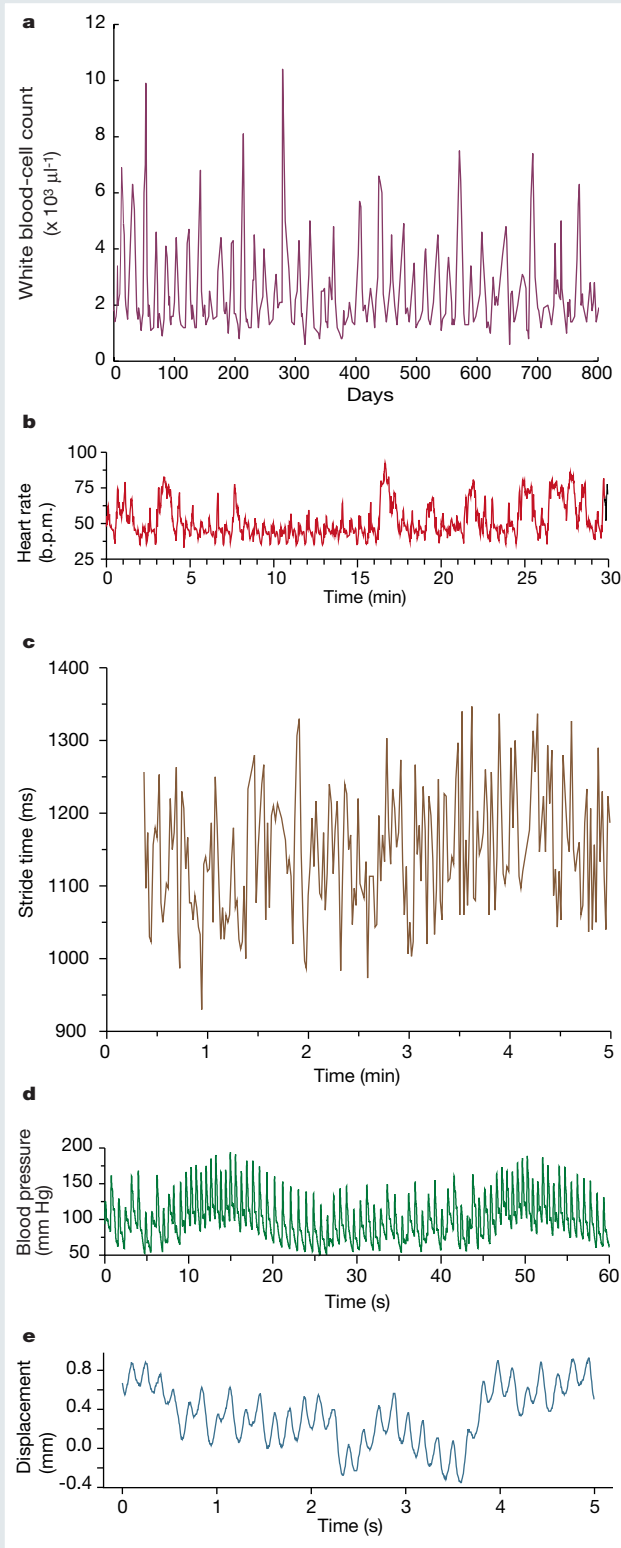


Figure 1 Representative physiological time series. **a**, White blood-cell count in a patient with cyclical neutropenia (tracing provided by D.C. Dale, C. Haurie and M. C. Mackey)³¹. **b**, Heart rate in a subject at high altitude (adapted with permission from the British Heart Journal)⁸⁵. **c**, Stride time in a patient with Huntington's disease (tracing provided by J. Hausdorff)^{86,87}. **d**, Blood pressure in a patient with sleep apnoea (tracing provided by A. Goldberger). **e**, Parkinsonian tremor measured from a finger (tracing provided by R. Edwards and A. Beuter⁷⁹). (The traces in panels **c** and **d** are adapted with permission from the Research Resource for Complex Physiologic Signals at <http://www.physionet.org>.)

Box 1

Properties of nonlinear equations

Nonlinear dynamics is a well developed discipline with many outstanding results⁸⁹. Three concepts are central — bifurcations, limit-cycle oscillations and chaos.

Bifurcations are changes in qualitative properties of dynamics. For example, as a parameter changes, steady states can become unstable and lead to stable oscillations, or a system with one stable steady state can be replaced by systems with multiple stable steady states. Physiological correlates are immediate. Drugs may lead to changes in control systems so that an abnormal, unhealthy rhythm is replaced by a more normal one. Mathematically, the drug induces a bifurcation in the dynamics, and as such, the actions of the drug can be analysed in a theoretical context. Often, the same type of bifurcation can be found in a host of different mathematical equations or experimental systems, and it is common to consider the 'universal' features of such bifurcations. Because many diseases are classified and identified by physicians based on characteristic qualitative features of dynamics, there is a natural match between the emphasis on qualitative dynamics in both mathematics and medicine.

Stable limit-cycle oscillations are a key feature of some nonlinear equations. Following a perturbation, a stable limit-cycle oscillation re-establishes itself with the same amplitude and frequency as before the perturbation. A perturbation to a linear oscillation may lead to a new amplitude of oscillation. For example, there is an intrinsic pacemaker that sets the rhythm in human hearts. If one or more electric shocks are delivered directly to the heart near the intrinsic pacemaker, the heart rhythm is modified transiently but re-establishes itself with the same frequency as before within a few seconds.

Chaos refers to aperiodic dynamics in deterministic equations in which there is a sensitivity to initial conditions. This means that even in the absence of stochastic processes, irregular rhythms can be generated. Although it is easy to consider mathematical systems in which all stochastic influences have been eliminated, in real physical and biological systems it is impossible to eliminate stochastic inputs. Thus, although chaotic dynamics is a clear mathematical concept, application of this concept to real biological systems is a difficult undertaking.

channels approaches the exponential distribution $e^{-0.1t}$. Although most models of channel opening and closing assume a stochastic mechanism such as the one above, deterministic chaotic models might also be consistent with the observed channel dynamics^{2,3}.

It is remarkable that the irregular openings and closings of ionic channels underlie all neural and muscular activities, including those that require great accuracy and precision such as reading this sentence, playing a violin concerto or carrying out a gymnastics routine. Because typical cells have of the order of at least 10^3 ionic channels of each type, deterministic equations can be used to model cells even though the underlying mechanism is probably stochastic^{11–13}. This notion is supported by studies of heart pacemaker cells in which channel dynamics have been modelled by random Markov processes. The resulting dynamics are similar to those generated by the deterministic models with an added 'noisy' fluctuation of period similar to what is observed experimentally^{14,15}. Additional regularization of dynamics can arise as a consequence of coupling of cells with irregular dynamics or cells that are heterogeneous^{15–17}.

Given the difficulty of analysing the origin of temporal fluctuations on a subcellular or cellular level, it is not surprising that the analysis of physiological rhythms in intact organisms provides added difficulties. In some cases, for example, for electrical activity of the

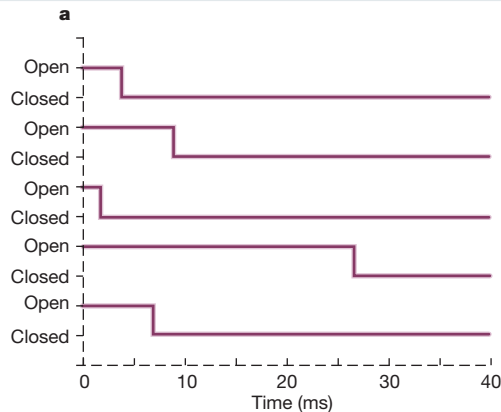
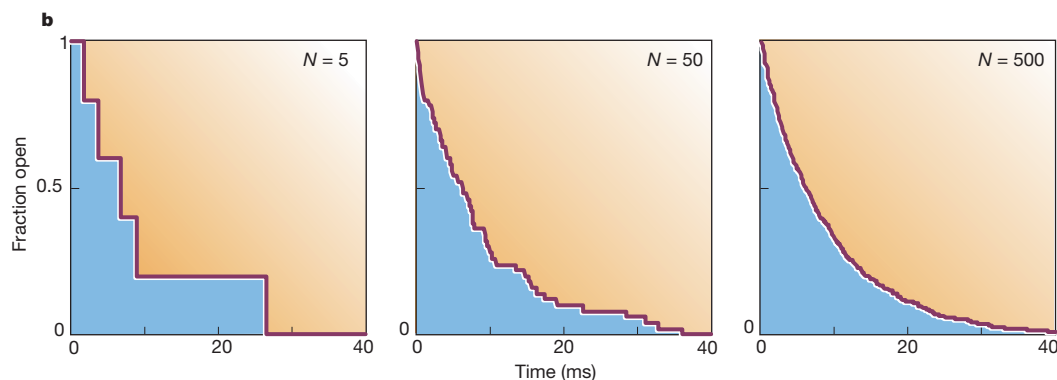


Figure 2 Schematic diagram showing dynamics in ionic channels that deactivate by a Poisson process (based on an analysis of dynamics observed in acetylcholine channels¹³). **a**, Channels that are open at $t = 0$, close randomly with a probability of 0.1 ms^{-1} . **b**, The fraction of open channels as a function of time for $N = 5$, $N = 50$ and $N = 500$ channels.



brain and the heart, the data can be collected easily using electrodes on the body surface, and computers provide a means for rapid analysis of large data sets. In other areas, such as endocrine function, determination of hormone fluctuations is difficult and expensive, requiring drawing blood at frequent intervals and performing expensive assays on the blood samples.

Consider the timing of the normal heartbeat as a paradigmatic example of physiological rhythm. Maintaining an adequate blood flow to the brain is essential to life. The heartbeat is generated by an autonomous pacemaker in the heart, but its frequency is mediated by neural activity controlled in turn by a large number of different feedback circuits all acting in parallel. As a consequence, the normal heart rate displays complex fluctuations in time in response to environmental factors such as breathing, exercise, changes in posture and emotion¹⁸. Diseases that impair heart function, for example damage to the heart caused by a heart attack or high blood pressure, lead to impaired pumping ability of the heart. In such cases, because the normal heart rate would not pump adequate blood to the brain, the heart generally beats more rapidly, and many of the normal fluctuations of heart rate are blunted. Lacking clearly defined and widely accepted operational definitions for chaotic dynamics, it is not surprising that agreement is lacking about whether or not normal heart dynamics are chaotic¹⁹ or not chaotic^{20,21}. Various tests of time variation have been applied to heart-rate variability to show that heart-rate fluctuation displays $1/f$ noise²², fractal dynamics with long-range correlations^{23,24}, and multifractal dynamics²⁵. Although similar dynamics in physical systems have been associated with self-organized criticality²⁶, in the biological context it is impossible to assert a mechanism based on the current observations. Indeed, it is possible that many of the properties observed reflect the response of individuals to a changing environment, and the environmental inputs themselves have interesting scaling properties²⁷.

There have been extensive quantitative analyses of data from many other physiological systems. In some instances the original motivation for the analyses was to determine whether the system dynamics was chaotic, although I believe the answer to this remains unclear. But

the data do reveal extraordinarily rich dynamics, and I mention several examples briefly. Electroencephalographic data reflect integrated activity from large numbers of cells in localized regions of the brain. A seizure shows characteristic changes in electroencephalographic records typified by larger-amplitude, regular and sustained oscillations reflecting broad synchronization of neural activity²⁸. Standard clinical interpretation of electroencephalographic data has been supplemented by a variety of quantitative analyses motivated by nonlinear dynamics^{29,30}. Analysis has also been carried out of a variety of normal and abnormal rhythms of blood-cell levels. Some blood disorders are characterized by pronounced fluctuations in levels of circulating white blood cells³¹. Blood-cell levels are controlled by feedback loops with long time delays and theoretical models have succeeded recently in analysing the effects of various manipulations. Endocrine function also provides a challenging area in view of the difficulty of obtaining temporal data, although sustained efforts have generated time series of various hormones including parathyroid hormone³², growth hormone³³ and prolactin³⁴.

These initial studies indicate extremely rich dynamics with differences between normal individuals and patients. The issue of whether or not the dynamics reflect chaos is much less interesting than elucidating the underlying mechanisms controlling the dynamics.

Synchronization of physiological oscillators

Although many cells in the body display intrinsic, spontaneous rhythmicity, physiological function derives from the interactions of these cells with each other and with external inputs to generate the rhythms essential for life. Thus, the heartbeat is generated by the sinoatrial node, a small region in the right atrium of the heart composed of thousands of pacemaker cells that interact with each other to set the normal cardiac rhythm^{14,15}. Nerve cells generating locomotion synchronize with definite phase relations depending on the species and the gait³⁵. And the intrinsic sleep–wake rhythm is usually synchronized to the light–dark cycle^{1,36}. In general, physiological oscillations can be synchronized to appropriate external or internal stimuli, so it is important to analyse the effects of

stimuli on intrinsic physiological rhythms. Even the simplest theoretical models (see Box 2 and Figs 3, 4) show the enormous complexity that can arise from periodic stimulation of nonlinear oscillations.

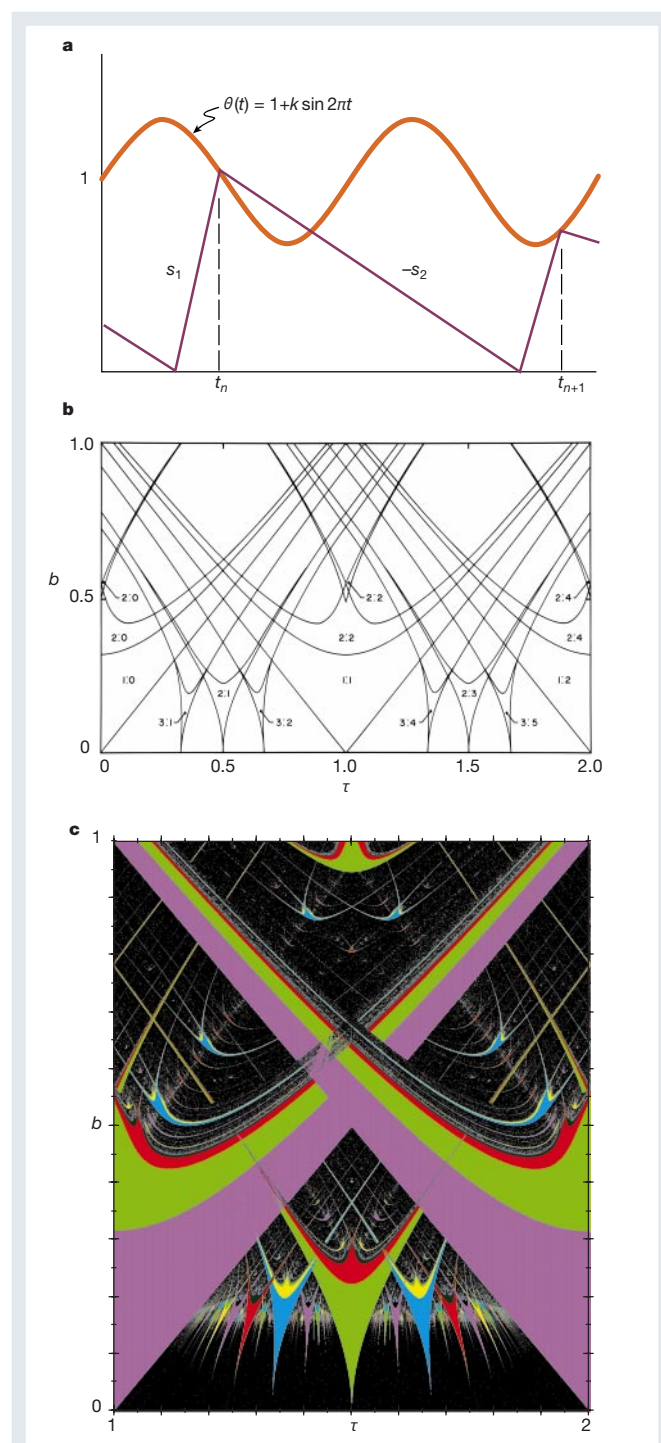


Figure 3 Integrate and fire model and locking zones. **a**, An activity increases linearly until it reaches a sinusoidal threshold and then decreases linearly. The successive times when the activity reaches threshold are determined iteratively. **b**, Schematic picture of the main Arnold tongues for the sine circle map (equation (3) in Box 2). $N:M$ locking reflects M cycles of the sine wave and N cycles of the activity. Chaotic dynamics are found only for $b > 1/2\pi$. **c**, Colour representation of a region of **b**. The colours code different periodic orbits (compare with **b**). The delicate geometry and self-similar structures are more evident in this representation. (Panels **a** and **b** modified with permission from ref. 88; colour version of the locking zones provided by J. Gallas.)

In vitro experiments in cardiac and neural tissue have clarified the effects of periodic stimulation in biological systems. Heart and nerve tissue are examples of excitable tissue⁴. This means that in response to a stimulus that is sufficiently large they will generate an event called an action potential. Following the action potential, for a time interval called the refractory period, a second stimulus does not elicit a second action potential. During periodic stimulation there are both periodic synchronized rhythms and aperiodic rhythms (see Box 2). Periodic stimulation of biological systems can give also give rise to aperiodic rhythms. For weak stimulation, it is common to find quasiperiodic rhythms, in which two rhythms with different frequencies march through each other with little interaction. Other aperiodic rhythms are chaotic. Identification of chaotic dynamics in periodically stimulated heart and nerve tissue is made more certain by the derivation of deterministic, theoretical models that are accurate both quantitatively and qualitatively and that correspond closely to both the regular and irregular dynamics observed experimentally^{6–8,10}. As the frequencies and amplitudes of the stimulation are varied, there is a characteristic organization of the phase-locking zones. However, if the dynamics are dissected on a fine scale, the detailed bifurcations in model systems differ from one another over some amplitude ranges. In experimental systems it is difficult to carry out locking studies in which stimulation parameters are varied finely because living preparations can be affected by stimulation and may change over the lengthy times needed to carry out the stimulation.

External inputs can often synchronize biological rhythms. Plants and animals display a circadian rhythm in which key processes show a 24-hour periodicity. This periodicity is usually set by the 24-hour light–dark cycle, but if this is removed by placing the organism in a constant environment, a cycle length different from 24 hours is observed. Thus the light–dark cycle entrains the intrinsic rhythm. If there is shift of the light–dark cycle, for example as might be generated by visiting a different time zone, then a time lag occurs until there is a new synchronization. Such phenomena have been modelled by both integrate and fire models and limit-cycle models^{36–39}.

Other circumstances in which physiological rhythms are stimulated by regular, periodic inputs occur in the context of medical devices. A mechanical ventilator assists breathing in experiments and in people who have respiratory failure. Such devices can be used in a variety of modes, but in the simplest, the physician sets the period and amplitude and the mix of gases for the ventilator, which then periodically inflates the patient's lungs. The resulting periodic lung inflation can interact with the person's intrinsic respiratory rhythm. In some instances, the respiratory rhythm will be entrained to the ventilator, but in other cases the patient will breath out when the ventilator is in the inflation phase^{40–42}.

These examples illustrate the effects of an external periodic input on intrinsic physiological rhythms. But the physiological rhythms also interact with one another. An example is the increase of the heart frequency during inspiration. Although the interactions between the respiratory and cardiac rhythms are not strong enough usually to lead to synchronization, such synchronization has been demonstrated in healthy high-performance swimmers⁴³.

It seems likely that many bodily activities require synchronization of cellular activity. For example, synchronization seems to be an essential component of many cortical tasks — coherent oscillations at 25–35 Hz are found in the sensorimotor cortex of monkeys⁴⁴ and 40–60-Hz oscillations are found in the visual cortex of cats⁴⁵.

Many different sorts of mathematical models have been proposed to account for synchronization in diverse systems. For example, synchronization has been observed in mathematical models of populations of cells that generate the heart beat in leeches¹⁶, the respiratory rhythm in rats⁴⁶, gamma and beta rhythms in the brain⁴⁷, and the circadian rhythm⁴⁸. However, because coupled oscillators show robust behaviour in which units tend to become synchronized⁴⁹, the observation of synchronization in models is not in itself a strong indicator of the suitability of the models.

Given the enormous numbers of connections between different cells and organs in the body, perhaps the biggest puzzle is not why bodily rhythms may in some circumstances become synchronized to

each other, but rather why there seems to be so much apparent independence between rhythms in different organs.

The function of noise and chaos in physiological systems

When looked at carefully, all physiological rhythms display fluctuations. Do the physiological systems function well despite these fluctuations, or are the fluctuations themselves intrinsic to the operation of the physiological systems? And if the fluctuations are essential to the functioning of the systems, is there some reason to believe that chaotic fluctuations would provide added function over stochastic fluctuations? There are no clear answers to these questions.

Chaotic dynamics might underlie normal function. In chaotic dynamics, there are typically many unstable periodic orbits, and there is also a sensitive dependence on initial conditions so that small, well-timed perturbations could lead to strong effects on the dynamics. Scientists have learned how to control chaotic dynamics⁵⁰ and nonlinear dynamics⁵¹ in model equations and in the laboratory. Electrical stimulation of complex dynamics in cardiac⁵² and neural⁵³ systems using chaos-control techniques has led to the regularization of complex rhythms, although the mechanisms of these effects are not certain⁵¹. Rabinovich and Abarbanel suggest that chaotic dynamics might be easier for the body to control than stochastic dynamics¹⁶. Another way the body might exploit chaos is by associating different unstable orbits with different percepts. Skarda and Freeman⁵⁴ proposed that the spatiotemporal fluctuations observed in the olfactory bulb in rabbits are chaotic. Each different odour is associated with and selects a different unstable spatiotemporal pattern of oscillation. However, in view of the difficulties associated with recording and interpreting spatiotemporal dynamics in living biological systems, and the huge gaps in understanding complex physiological functions involved in cognition and control, these claims remain intriguing hypotheses rather than established fact.

In normal circumstances, detection of signals is hampered by noise. For example, because the aesthetic and practical utility of sounds and images is reduced as noise level increases, designers of devices for recording and playback of sound and images make strenuous efforts to maximize the signal-to-noise ratio.

In other circumstances, however, the presence of noise and/or chaotic dynamics can improve detection of signals. Stochastic resonance refers to a situation in which the signal-to-noise ratio is maximum at an intermediate level of noise^{55–58}. For example, in tasks that are at the threshold of perception, the addition of noise can improve threshold detection. To illustrate this, consider a 'leaky' integrate and fire model⁵⁹ in the presence of noise as a model for stochastic resonance (Fig. 5). Assume that an activity (here a membrane potential of a nerve cell) x is governed by the differential equation

$$\frac{dx}{dt} = -cx + I + \xi \quad (1)$$

where c is a decay constant, I is constant input and ξ is added random noise. The activity rises to a threshold $f(t) = 1 + k \sin 2\pi t$, which represents an oscillating signal. If $c = 0$, then x increases linearly to the threshold (as shown in Fig. 3 and Box 2). When the activity reaches the threshold it elicits an action potential indicated by the arrows, followed by an immediate reset of the activity to zero. With $\xi = 0$ the maximum value of x is I/c . Consequently, if $(1 - k) > I/c$ and there is very low noise, then the activity would never reach threshold (Fig. 5a). But if there is moderate noise the activity can reach threshold, and this tends to occur at the troughs of the sine wave (Fig. 5b). Of course, if the noise is too great, then the activity is no longer as concentrated at the troughs (Fig. 5c). Thus, at intermediate values of the noise, the activity is well synchronized to the sinusoidal input, but rather than occurring on every cycle, there is a stochastic element resulting in a skipping pattern of activity in which excitation tends to occur at random multiples of a common divisor. Although the mechanisms are not well understood, similar rhythms, which may represent a 'stochastic phase locking', are observed frequently in physiological data^{56,57}.

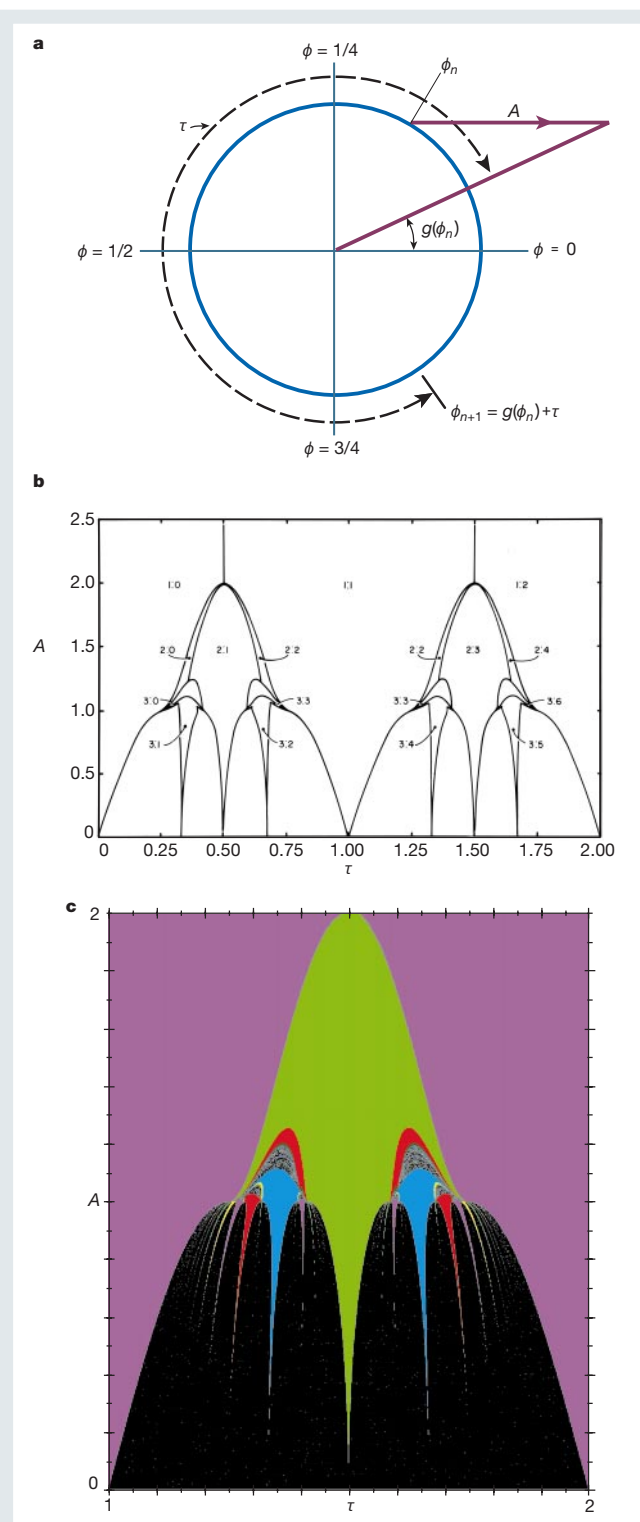


Figure 4 Poincaré oscillator model and locking zones. **a**, A stimulus of amplitude A delivered to the limit-cycle oscillator in equation (4) (Box 2) resets the phase. **b**, Schematic picture of the main Arnold tongues for the periodically forced Poincaré oscillator (equations (4) and (5) in Box 2). N/M locking reflects N stimuli for M cycles of the limit cycle oscillation. The Arnold tongue structure is present for $0 \leq A \leq 1$. **c**, Colour representation of a region of **b**. (Panels **a** and **b** modified with permission from ref. 88; colour version of the locking zones provided by J. Gallas.)

Box 2

Simple models for synchronization of nonlinear oscillators

Because biological oscillators are stable nonlinear oscillations, many of the general features of the interactions between the periodic input and the ongoing rhythm can be predicted without knowing precise details of the nature of the oscillator or the input. Simple toy models of synchronization show a host of phenomena shared by more complex models and real systems^{2,4,49,59,88,90–92}.

The integrate and fire model assumes that there is an activity that rises to a periodically oscillating threshold^{59,90} and then resets to a lower threshold. Depending on the setting, the activity might represent a nerve membrane potential, a substance that induces sleep or a neural activity associated with the timing of inspiration. There are also many ways in which one might model the activity. It could be modelled by a function that increases linearly in time, or by a function that saturates, or perhaps even by a biased random walk. One of the simplest embodiments of the integrate and fire model is to assume an oscillatory threshold $\theta(t) = 1 + k \sin 2\pi t$ and an activity that increases linearly to the threshold, followed by a linear decrease to zero (Fig. 3a). To illustrate some of the properties of this model, consider the situation in which there is a jump from 0 to the threshold, followed by a linear decay (slope of $-s_2$) to zero. Let t_n be the time when the oscillation is at the threshold value for the n th time. Then

$$t_{n+1} = t_n + 1/s_2 + (k/s_2) \sin 2\pi t_n \quad (2)$$

If $\tau = 1/s_2$, $b = k/s_2$ and $\phi_n = t_n \pmod{1}$, we obtain

$$\phi_{n+1} = \phi_n + \tau + b \sin 2\pi \phi_n \pmod{1} \quad (3)$$

This difference equation — sometimes called the sine circle map — displays an incredible array of complex rhythms⁹¹. One type of rhythm is synchronization of the activity to the sine function in an $N:M$ phase-locked rhythm so that for each M cycles of the sine wave the activity reaches the threshold N times. For low amplitudes ($0 \leq b \leq 1/2\pi$) of the sine-wave modulation, there is an orderly arrangement of the phase-locking zones, called Arnold tongues. For fixed values of b , as τ increases all rational ratios M/N are observed and there are quasiperiodic rhythms in which the two rhythms march through each other with slight interaction. For $b > 1/2\pi$, the simple geometry of Arnold tongues breaks down and there are chaotic dynamics as well as bistability in which two different stable rhythms exist simultaneously. Figure 3b,c gives a hint at the complexity of the organization of the locking zones.

A second prototypical oscillator was described originally by the French mathematician Poincaré

$$\frac{dr}{dt} = cr(1-r), \quad \frac{d\phi}{dt} = 2\pi \quad (4)$$

where the equation is written in a radial coordinate system (r, ϕ) . In this equation, there is a stable limit cycle at $r = 1$. More realistic models of biological oscillations also have limit cycles, but the variables in these models would reflect the mechanism of the underlying biological oscillation. For example, more realistic models of biological oscillators might be written in terms of ionic currents or neural activities. Because there are certain features of the qualitative response of biological oscillators to single and periodic stimulation that depend solely on the properties of nonlinear oscillations¹, the tractable analytic properties of the Poincaré oscillator make it an ideal system for theoretical analysis. We assume that the stimulus is a translation by a horizontal distance A from the current state point, followed by evolution according to equation (4) (Fig. 4a). If c is sufficiently large, then after the periodic perturbation there is a very rapid return to the limit cycle, and the dynamics can be described approximately by⁹²

$$\phi_{n+1} = \tau + g(\phi_n, A) \pmod{1}, \quad \text{where } g(\phi, A) = \frac{1}{2\pi} \arccos \frac{\cos 2\pi \phi + A}{(1 + A^2 + 2A \cos 2\pi \phi)^{1/2}} \quad (5)$$

Although for $0 \leq A \leq 1$ there is again a simple Arnold-tongue geometry that is the same topologically as that found in the integrate and fire model for low amplitudes, for higher amplitudes the detailed geometry is radically different (Fig. 4b, c).

Although noise is helpful in this artificial example, various proposals have been made suggesting that signal detection might be facilitated in physiological systems by similar mechanisms in which noise was either added externally or present intrinsically. Several reports indicate that added noise seems to enhance signal detection in *in vitro* preparations⁵⁸, in animal experiments^{60,61} and in human perception^{62–64}. Because noise is an intrinsic property of human physiological systems, a compelling question is whether or not noise may be acting to aid function in normal activities in a manner similar to the way it can act to enhance the detection of subthreshold activities.

Prospects for medical applications

In the physical sciences, advances in basic science have inevitably led to new technology. Although many technological innovations in medicine, such as X-ray imaging and nuclear magnetic resonance imaging, have derived from advances in the physical sciences, the recent mathematical analyses of temporal properties of physiological

rhythms have not yet led to medical advances, although several directions are under active consideration.

There is a wide spectrum of dynamical behaviour associated with both normal and pathological physiological functioning. Extremely regular dynamics are often associated with disease, including periodic (Cheyne–Stokes) breathing, certain abnormally rapid heart rhythms, cyclical blood diseases, epilepsy, neurological tics and tremors. However, regular periodicity can also reflect healthy dynamics — for example in the sleep–wake cycle and menstrual rhythms. Finally, irregular rhythms can also reflect disease. Cardiac arrhythmias such as atrial fibrillation and frequent ectopy, and neurological disorders such as post-anoxic myoclonus, are often highly irregular. The term ‘dynamical disease’ captures the notion that abnormal rhythms, which could be either more irregular or more regular than normal, arise owing to modifications in physiological control systems that lead to bifurcations in the dynamics⁶⁵. What is most important in distinguishing health from disease is that

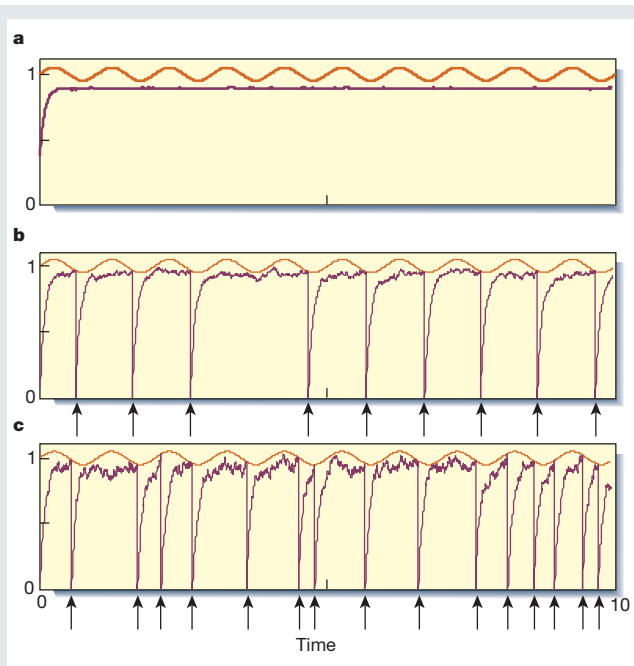


Figure 5 Schematic representation of a noisy, leaky integrate and fire model with three levels of noise. The activity rises to a sinusoidally modulated threshold following equation (1). **a**, The activity saturates and does not reach threshold. **b, c**, As the noise increases, the activity reaches threshold leading to an action potential (arrows). The ability to detect the frequency of the sine wave by the induced spiking activity will have a maximum efficiency at some intermediate level of noise. This illustrates the phenomenon of stochastic resonance^{55,57}.

there is a change in the dynamics from what is normal, rather than regularity or irregularity of the dynamics. Thus, generalizations such as “chaos may provide a healthy flexibility to the heart, brain, and other parts of the body”⁷⁶ must be considered warily.

Efforts are underway to develop better diagnostic and prognostic methods by analysis of dynamics of physiological rhythms. For example, in cardiology it would be extremely useful to have a measure that provides an independent predictor of significant cardiac events such as impending cardiac arrest. A variety of quantitative measures of cardiac activity derived from concepts introduced in nonlinear dynamics have been proposed as markers of increased risk of sudden death^{67–71}. However, because algorithms must be tested on large populations, it is difficult to design and implement clinical trials to test the utility of the various proposed measures. One recent initiative, the development of open databases that could serve as a repository of clinical data that are difficult and expensive to obtain, provides an emerging strategy that should prove indispensable for testing competing algorithms⁷². Likewise, the public availability of powerful data-processing methods should advance research⁷³.

Similar efforts at analysis of dynamics of time series are also underway in neurology. The earlier debates about whether the electroencephalogram displayed chaotic dynamics have been superseded by predictions of the onset of seizures using a variety of measures derived from nonlinear dynamics^{74–76}. Analysis of abnormalities in tremor and standing steadiness may provide a marker of heavy-metal toxicity^{77,78} or the early onset of Parkinson’s disease⁷⁹.

There are a number of ways in which knowledge about the synchronization of oscillators could be put to medical use. Because bodily functions show distinct periodicities, schedules for drug administration might be optimized. In cancer chemotherapy, treatments could be based on the circadian rhythm of cell division⁸⁰. Intriguing strategies to minimize the effects of jet lag have been developed, based on experimental studies of resetting the circadian

oscillator by modifying the exposure to light after travel^{1,81}, although I am unaware of any clinical studies that have assessed these proposals.

Medical devices that are used to regulate and artificially control cardiac and respiratory dynamics have been developed principally by engineers working with physicians. The empirical methods used to develop and test these devices have not included a detailed mathematical analysis of interactions of the physiological rhythm with the device. Indeed, devices usually have several adjustable parameters so that the physician can optimize the settings for operating the device based on direct testing in a patient. Recent work has investigated the application of algorithms motivated by nonlinear dynamics to control cardiac arrhythmias in humans (ref. 82 and D. J. Christini, K. M. Stein, S. M. Markowitz, S. Mittal, D. J. Slotwimer, M. A. Scheiner, S. Iwai and B. B. Lerman, unpublished results). Although there are not yet clinical applications, I anticipate that better understanding of the interactions between stimuli and physiological rhythms will lead to the development of better medical devices.

The recent identification of stochastic resonance in experimental systems has led to the suggestion that it might be useful to add noise artificially as a therapeutic measure. For example, in patients who have suffered strokes or peripheral nerve damage, detection tasks might be enhanced by addition of artificial noise to enhance tactile sensations⁶². Similarly, addition of sub-sensory mechanical noise applied to the soles of the feet of quietly standing healthy subjects reduced postural sway and seemed to stabilize the postural control (J. Niemi, A. Priplata, M. Salen, J. Harry and J. J. Collins, unpublished results). Another intriguing suggestion is that addition of noise might also enhance the efficiency of mechanical ventilators. The use of a variable inflation volume with occasional large inflations might act to prevent the collapse of alveoli and therefore maintain improved lung function in patients undergoing ventilation⁸³. Finally, low-level vibratory stimulation of a premature infant seemed to eliminate long apnoeic periods. Although the mechanism is unknown, it was hypothesized that a stable fixed point, corresponding to no breathing in the apnoeic infant, coexisted with the normal limit-cycle oscillation that corresponded to breathing. The vibration destabilized that fixed point thereby eliminating the apnoeic spells⁸⁴.

Conclusions

Physiological rhythms are generated by nonlinear dynamical systems. *In vitro* experimental systems often yield dynamics that can be successfully interpreted using both simplified and complicated mathematical models. These models make predictions about effects induced by parameter changes such as changing the frequency and amplitude of a periodic stimulus. However, rhythms in intact animals in an ambient environment have so far defied simple interpretation. Bodily rhythms such as the heart beat, respiration and cortical rhythms show complex dynamics, the function and origin of which are still poorly understood. Moreover, we do not understand if the complex dynamics are an essential feature, or if they are secondary to internal feedback and environmental fluctuations. Because of the complexity of biological systems and the huge jump in scale from a single ionic channel to the cell to the organ to the organism, for the foreseeable future all computer models will be gross approximations to the real system. In the physical sciences, scientific understanding has been expressed in elegant theoretical constructs and has led to revolutionary technological innovation. If the advances in understanding physiological rhythms will follow the same trajectory, then we are still just at the beginning. □

1. Winfree, A. T. *The Geometry of Biological Time* (Springer, New York, 1980; 2nd edn 2001).
2. Glass, L. & Mackey, M. C. *From Clocks to Chaos: The Rhythms of Life* (Princeton Univ. Press, Princeton, 1988).
3. Bassingthwaite, J. B., Liebovitch, L. S. & West, B. J. *Fractal Physiology* (Oxford Univ. Press, New York, 1994).
4. Keener, J. & Sneyd, J. *Mathematical Physiology* (Springer, New York, 1998).
5. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol. (Lond.)* **117**, 500–544 (1952).
6. Aihara, K. & Matsumoto, G. in *Chaos in Biological Systems* (eds Degn, H., Holden, A. V. & Olson, L. F.) 121–131 (Plenum, New York, 1987).

7. Takahashi, N., Hanyu, Y., Musha, T., Kubo, R. & Matsumoto, G. Global bifurcation structure in periodically stimulated giant axons of squid. *Physica D* **43**, 318–334 (1990).
8. Kaplan, D. T. *et al.* Subthreshold dynamics in periodically stimulated squid giant axons. *Phys. Rev. Lett.* **76**, 4074–4077 (1996).
9. Ruelle, D. Where can one hope to profitably apply the ideas of chaos? *Phys. Today* **47**, 29–30 (July 1994).
10. Guevara, M. R., Glass, L. & Shrier, A. Phase-locking, period-doubling bifurcations and irregular dynamics in periodically stimulated cardiac cells. *Science* **214**, 1350–1353 (1981).
11. DeFelice, L. J. & Clay, J. R. in *Single-Channel Recording* Ch. 15 (eds Sakmann, B. & Neher, E.) 323–342 (Plenum, New York, 1983).
12. DeFelice, L. J. & Isaac, A. Chaotic states in a random world: relationships between the nonlinear differential equations of excitability and the stochastic properties of ion channels. *J. Stat. Phys.* **70**, 339–354 (1993).
13. Colquhoun, D. & Hawkes, A. G. in *Single-Channel Recording* Ch. 20 (eds Sakmann, B. & Neher, E.) 397–482 (Plenum, New York, 1995).
14. Wilders, R. & Jongsma, H. J. Beating irregularity of single pacemaker cells isolated from the rabbit sinoatrial node. *Biophys. J.* **60**, 2601–2613 (1993).
15. Guevara, M. R. & Lewis, T. A minimal single channel model for the regularity of beating of the sinoatrial node. *Chaos* **5**, 174–183 (1995).
16. Rabinovich, M. I. & Abarbanel, H. D. I. The role of chaos in neural systems. *Neuroscience* **87**, 5–14 (1998).
17. De Vries, G., Sherman, A. & Zhu, H. R. Diffusively coupled bursters: effects of cell heterogeneity. *B. Math. Biol.* **60**, 1167–1200 (1998).
18. Camm, A. J. *et al.* Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation* **93**, 1043–1065 (1996).
19. Poon, C.-S. & Merrill, C. K. Decrease of cardiac chaos in congestive heart failure. *Nature* **389**, 492–495 (1998).
20. Kanter, J. K., Holstein-Rathlou, N.-H. & Agner, E. Lack of evidence for low-dimensional chaos in heart rate variability. *J. Cardiovasc. Electrophysiol.* **5**, 591–601 (1994).
21. Costa, M. *et al.* No evidence of chaos in the heart rate variability of normal and cardiac transplant human subjects. *J. Cardiovasc. Electrophysiol.* **10**, 1350–1357 (1999).
22. Kobayashi, M. & Musha, T. I/f fluctuation of heartbeat period. *IEEE Trans. Biomed. Eng.* **29**, 456–457 (1982).
23. Peng, C. K. *et al.* Long-range anticorrelations and non-Gaussian behavior of the heartbeat. *Phys. Rev. Lett.* **70**, 1343–1346 (1993).
24. Turcott, R. G. & Teich, M. Fractal character of the electrocardiogram: distinguishing heart failure and normal patients. *Ann. Biomed. Eng.* **24**, 269–293 (1996).
25. Ivanov, P. C. *et al.* Multifractality in human heartbeat dynamics. *Nature* **399**, 461–465 (1999).
26. Bak, P. *How Nature Works: The Science of Self-Organized Criticality* (Copernicus, New York, 1996).
27. Roach, D., Sheldon, A., Wilson, W. & Sheldon, R. Temporally localized contributions to measures of large-scale heart rate variability. *Am. J. Physiol.* **274** (Heart Circ. Physiol. H43), H1465–H1471 (1998).
28. Kiloh, L. *et al.* *Clinical Electroencephalography* (Butterworths, London, 1981).
29. Babloyantz, A. & Destexhe, A. Low-dimensional chaos in an instance of epilepsy. *Proc. Natl Acad. Sci. USA* **83**, 3513–3517 (1986).
30. Jansen, B. H. & Brandt, M. E. (eds) *Nonlinear Dynamical Analysis of the EEG* (World Scientific, Singapore, 1993).
31. Haurie, C., Dale, D. C. & Mackey, M. C. Cyclical neutropenia and other periodic hematological disorders: a review of mechanisms and mathematical models. *Blood* **92**, 2629–2640 (1998).
32. Prank, K. *et al.* Nonlinear dynamics in pulsatile secretion of parathyroid hormone in human subjects. *Chaos* **5**, 76–81 (1995).
33. Prank, K. *et al.* Self-organized segmentation of time series: separating growth hormone secretion in acromegaly from normal controls. *Biophys. J.* **70**, 2540–2547 (1996).
34. Veldman, R. G. *et al.* Increased episodic release and disorderliness of prolactin secretion in both micro- and macroprolactinomas. *Eur. J. Endocrinol.* **140**, 192–200 (1999).
35. Golubitsky, M., Stewart, I., Buono, P. L. & Collins, J. J. Symmetry in locomotor central pattern generators and animal gaits. *Nature* **401**, 693–695 (1999).
36. Strogatz, S. H. *The Mathematical Structure of the Human Sleep-Wake Cycle. Lecture Notes in Biomathematics* Vol. 69 (Springer, Berlin, 1986).
37. Daan, S., Beersma, D. G. M. & Borbely, A. A. Timing of human sleep: recovery process gated by a circadian pacemaker. *Am. J. Physiol.* **246**, R161–R178 (1984).
38. Jewett, M. E. & Kronauer, R. E. Refinement of a limit cycle oscillator model of the effects of light on the human circadian pacemaker. *J. Theor. Biol.* **192**, 455–465 (1998).
39. Leloup, J. C. & Goldbeter, A. A model for circadian rhythms in *Drosophila* incorporating the formation of a complex between the PER and TIM proteins. *J. Biol. Rhythms* **13**, 70–87 (1998).
40. Petriello, G. A. & Glass, L. A theory for phase-locking of respiration in cats to a mechanical ventilator. *Am. J. Physiol.* **246** (Regulat. Integrat. Comp. Physiol. 15), R311–R320 (1984).
41. Graves, C., Glass, L., Laporta, D., Meloche, R. & Grassino, A. Respiratory phase-locking during mechanical ventilation in anesthetized human subjects. *Am. J. Physiol.* **250** (Regulat. Integrat. Comp. Physiol. 19), R902–R909 (1986).
42. Simoin, P. M., Habel, A. M., Daubenspeck, J. A. & Leiter, J. C. Vagal feedback in the entrainment of respiration to mechanical ventilation in sleeping humans. *J. Appl. Physiol.* **89**, 760–769 (2000).
43. Schäfer, C., Rosenblum, M. G., Kurths, J. & Abel, H.-H. Heartbeat synchronized with ventilation. *Nature* **392**, 239–240 (1998).
44. Murthy, V. N. & Fetz, E. E. Coherent 25- to 35-Hz oscillations in the sensorimotor cortex of awake behaving monkeys. *Proc. Natl Acad. Sci. USA* **89**, 5670–5674 (1992).
45. Gray, C. M., König, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334–337 (1989).
46. Butera, R. J., Rinzel, J. & Smith, J. C. Models of respiratory rhythm generation in the pre-Bötzinger complex. II. Populations of coupled pacemaker neurons. *J. Neurophysiol.* **82**, 398–415 (1999).
47. Kopell, N., Ermentrout, G. B., Whittington, M. A. & Traub, R. D. Gamma rhythms and beta rhythms have different synchronization properties. *Proc. Natl Acad. Sci. USA* **97**, 1867–1872 (2000).
48. Achermann, P. & Kunz, H. Modeling circadian rhythm generation in the suprachiasmatic nucleus with locally coupled self-sustained oscillators: phase shifts and phase response curves. *J. Biol. Rhythms* **14**, 460–468 (1999).
49. Mirollo, R. E. & Strogatz, S. H. Synchronization properties of pulse-coupled biological oscillators. *SIAM J. Appl. Math.* **50**, 1645–1662 (1990).
50. Shinbrot, T. Progress in the control of chaos. *Adv. Phys.* **95**, 73–111 (1995).
51. Christini, D. J., Hall, K., Collins, J. J. & Glass, L. in *Handbook of Biological Physics* Vol. 4: *Neuroinformatics, Neural Modelling* (eds Moss, F. & Gielen, S.) 205–227 (Elsevier, Amsterdam, 2000).
52. Garfinkel, A., Spano, M. L., Ditto, W. L. & Weiss, J. N. Experimental control of cardiac chaos. *Science* **257**, 1230–1235 (1992).
53. Schiff, S. J. *et al.* Controlling chaos in the brain. *Nature* **370**, 615–620 (1994).
54. Skarda, C. A. & Freeman, W. J. How brains make chaos in order to make sense of the world. *Behav. Brain Sci.* **10**, 161–195 (1987).
55. Gammaitoni, L., Hänggi, P., Jung, P. & Marchesoni, F. Stochastic resonance. *Rev. Mod. Phys.* **70**, 223–288 (1998).
56. Longtin, A., Bulsara, A. & Moss, F. Time-interval sequences in bistable systems and the noise-induced transmission of information by sensory neurons. *Phys. Rev. Lett.* **67**, 656–659 (1991).
57. Longtin, A. Mechanisms of stochastic phase locking. *Chaos* **5**, 209–215 (1995).
58. Douglass, J. K., Wilkens, L., Pantazidou, E. & Moss, F. Stochastic resonance: noise-enhanced information transfer in crayfish mechanoreceptors. *Nature* **365**, 337 (1993).
59. Keener, J. P., Hoppensteadt, F. C. & Rinzel, J. Integrate-and-fire models of nerve membrane response to oscillatory input. *SIAM J. Appl. Math.* **41**, 503–517 (1981).
60. Collins, J. J., Imhoff, T. T. & Grigg, P. Noise-enhanced information transmission in rat SA1 cutaneous mechanoreceptors via aperiodic stochastic resonance. *J. Neurophysiol.* **76**, 642–645 (1996).
61. Greenwood, P. E., Ward, L. M., Russell, D. F., Neiman, A. & Moss, F. Stochastic resonance enhances the electrosensory information available to paddlefish for prey capture. *Phys. Rev. Lett.* **84**, 4773–4776 (2000).
62. Collins, J. J., Imhoff, T. T. & Grigg, P. Noise-mediated enhancements and decrements in human tactile sensation. *Phys. Rev. E* **56**, 923–926 (1997).
63. Simonotto, E. *et al.* Visual perception of stochastic resonance. *Phys. Rev. Lett.* **78**, 1186–1189 (1997).
64. Richardson, K. A., Imhoff, T. T., Grigg, P. & Collins, J. J. Using electrical noise to enhance the ability of humans to detect subthreshold mechanical stimuli. *Chaos* **8**, 599–603 (1998).
65. Mackey, M. C. & Glass, L. Oscillation and chaos in physiological control systems. *Science* **197**, 287–289 (1977).
66. Pool, R. Is it healthy to be chaotic? *Science* **243**, 604–607 (1989).
67. Skinner, J. E., Pratt, C. M. & Vybiral, T. A reduction in the correlation dimension of heartbeat intervals precedes imminent ventricular fibrillation in human subjects. *Am. Heart J.* **125**, 731–743 (1993).
68. Rosenbaum, D. S. *et al.* Electrical alternans and vulnerability to ventricular arrhythmias. *N. Engl. J. Med.* **330**, 235–241 (1994).
69. Huikuri, H. V. *et al.* Fractal correlation properties of R-R interval dynamics and mortality in patients with depressed left ventricular function after an acute myocardial infarction. *Circulation* **101**, 47–53 (2000).
70. Klingenstein, T. *et al.* Predictive value of T-wave alternans for arrhythmic events in patients with congestive heart failure. *Lancet* **356**, 651–652 (2000).
71. Lipsitz, L. A. Age-related changes in the “complexity” of cardiovascular dynamics: a potential marker of vulnerability to disease. *Chaos* **5**, 102–109 (1995).
72. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220 (2000). [Circulation Electronic Pages <<http://circ.ahajournals.org/cgi/content/abstract/101/23/e215>> (13 June 2000); see also <<http://www.physionet.org>>.]
73. Hegger, R., Kantz, H. & Schreiber, T. Practical implementation of nonlinear time series methods: the TISEAN package. *Chaos* **9**, 413–435 (1999); <<http://www.mpiikps-dresden.mpg.de/~tisean/>>.
74. Schiff, S. J. Forecasting brain storms. *Nature Med.* **4**, 1117–1118 (1998).
75. Lehnertz, K. & Elger, C. E. Can epileptic seizures be predicted? Evidence from nonlinear time series analysis of brain activity. *Phys. Rev. Lett.* **80**, 5019–5022 (1998).
76. Martinier, J. *et al.* Epileptic seizures can be anticipated by non-linear analysis. *Nature Med.* **4**, 1173–1176 (1998).
77. Beuter, A. & Edwards, R. Tremor in Cree subjects exposed to methylmercury: a preliminary study. *Neurotoxicol. Teratol.* **20**, 581–589 (1998).
78. Gerr, F., Letz, R. & Green, R. C. Relationships between quantitative measures and neurologist’s clinical rating of tremor and standing steadiness in two epidemiological studies. *NeuroToxicology* **21**, 753–760 (2000).
79. Edwards, R. & Beuter, A. Using time domain characteristics to discriminate physiologic and parkinsonian tremors. *J. Clin. Neurophysiol.* **17**, 87–100 (2000).
80. Mormont, M. C. & Lévi, F. Circadian-system alterations during cancer processes: a review. *Int. J. Cancer* **70**, 241–247 (1997).
81. Daan, S. & Lewy, A. J. Scheduled exposure to daylight: a potential strategy to reduce jet lag following transmeridian flight. *Psychopharmacol. Bull.* **20**, 566–588 (1984).
82. Ditto, W. L. *et al.* Control of human atrial fibrillation. *Int. J. Bif. Chaos* **10**, 593–601 (2000).
83. Suki, B. *et al.* Life-support systems benefits from noise. *Nature* **393**, 127–128 (1998).
84. Paydarfar, D. & Buerkle, D. M. Sporadic apnea: paradoxical transformation to eupnea by perturbations that inhibit inspiration. *Med. Hypotheses* **49**, 19–26 (1997).
85. Lipsitz, L. A. *et al.* Heart rate and respiratory dynamics on ascent to high altitude. *Br. Heart J.* **74**, 390–396 (1995).
86. Hausdorff, J. M. *et al.* Altered fractal dynamics of gait: reduced stride-interval correlations with aging and Huntington’s disease. *J. Appl. Physiol.* **82**, 262–269 (1997).
87. Hausdorff, J. M. *et al.* Dynamic markers of altered gait rhythm in amyotrophic lateral sclerosis. *J. Appl. Physiol.* **88**, 2045–2053 (2000).
88. Glass, L. Cardiac arrhythmias and circle maps—a classical problem. *Chaos* **1**, 13–19 (1991).
89. Wiggins, S. *Introduction to Applied Nonlinear Dynamical Systems and Chaos* (Springer, New York, 1990).
90. Glass, L. & Mackey, M. C. A simple model for phase locking of biological oscillators. *J. Math. Biol.* **7**, 339–352 (1979).
91. Perez, R. & Glass, L. Bistability, period doubling bifurcations and chaos in a periodically forced oscillator. *Phys. Lett.* **90A**, 441–443 (1982).
92. Guevara, M. R. & Glass, L. Phase-locking, period-doubling bifurcations and chaos in a mathematical model of a periodically driven biological oscillator: a theory for the entrainment of biological oscillators and the generation of cardiac dysrhythmias. *J. Math. Biol.* **14**, 1–23 (1982).

Acknowledgements

Thanks to M. R. Guevara, A. L. Goldberger, J. J. Collins, J. Milton and E. Cooper for helpful conversations; J. Lacuna, Y. Nagai and T. Inoue for assistance with the figures; and J. Gallas (Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil) for providing the colour representations of the locking zones in Figs 3c and 4c. My research has been supported by NSERC, MRC, MITACS, Canadian Heart and Stroke Foundation, FCAR and the Research Resource for Complex Physiologic Signals (NIH).

From phage to page

Phage-resistant cells, dsDNA markers and plenty of SNP data on the web.

Amplifluor primers

Intergen www.intergen.co.uk

Universal hairpin primers

Amplifluor primers use a 5'-hairpin structure and a pair of energy-transfer molecules (a fluorophore and a quencher) to generate a fluorescence signal only when incorporated into the amplicon. Intergen now offers a method of SNP (single-nucleotide polymorphism) discrimination using multiplex PCR with universal hairpin primers. Two primers that discriminate the SNP at the 3'-terminal base are tailed on their 5'-ends with two different universal sequences. These tails are recognized by the two hairpin primers, one labelled with fluorescein (FAM) and the other with sulphorhodamine (SR). A common reverse primer is shared by both PCR primer sets. Depending on which base is present in the sequence, either the FAM or the SR signal is greater. Optimized for both real-time and end-point PCR, the system is designed for high-throughput analyses.

Reader Service No. 100

RNeasy Protect

Qiagen www.qiagen.com

A complete RNA protection and isolation solution in one kit

The RNeasy Protect system combines RNeasy technology with the RNAlater RNA stabilization reagent for purification of high-quality, intact RNA from a range of biological materials. A problem when handling RNA is its extreme instability following harvesting of samples. The aqueous Qiagen RNAlater reagent provides RNA stabilization and protection *in situ*, it rapidly permeates through tissue to ensure subsequent reliable gene-expression and gene-profiling data. There is no need for liquid-nitrogen or dry-ice processing. Even after multiple freeze-thaw cycles biological samples can be successfully used and/or archived without the risk of RNA degradation. Following stabilization, RNeasy silica-gel-membrane technology delivers RNA ready-to-use for downstream processes such as northern blotting and RT-PCR. The kits are available in a choice of mini, midi and maxi formats for the purification of up to 100 µg, 1 mg or 6 mg RNA respectively.

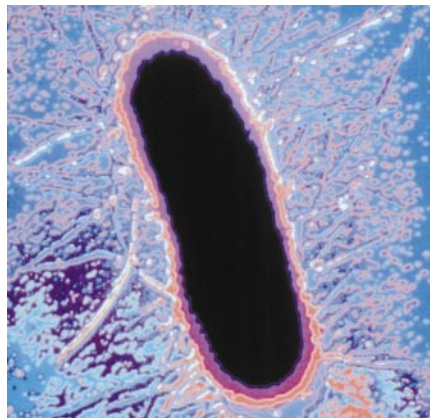
Reader Service No. 101

T1 resistance

Life Technologies www.lifetech.com

T1 phage-resistant competent cells

Researchers needing to store clones without fear of loss, and to protect cDNA, libraries can



Phage fighter: clone protection by GIBCO BRL.

now choose between two strains of competent cells resistant to T1, the bacteriophage that infects and destroys *E. coli*. GIBCO BRL MAX Efficiency DH5a and ElectroMAX DH10B are available from Life Technologies. Efficiency and performance of the D and DH10B are said to be equal to that of non-T1 phage-resistant cells, yielding 1×10^9 c.f.u. per mg DNA.

Reader Service No. 102

MAb clone M86

Alexis www.alexis-corp.com

IgM isotype for xenotransplantation studies

Monoclonal antibody clone M86 is suitable for measuring α -Gal epitope expression on cells, glycolipids and glycoproteins and characterizing hyper-acute rejection (HAR) in organ and tissue transplantation. It can also be used to conduct cytotoxicity assays for α -Gal-specific pathways, monitor xenotransplantation experiments, and screen inhibitors of antibody-induced cytotoxicity. An IgM isotype that mimics *in vivo* reaction, Mab M86 is highly specific for synthetic and naturally produced Gal α 1,3 Gal epitopes on glycoproteins and glycolipids and exhibits broad species crossreactivity. It does not crossreact with β -Gal glycoproteins or BSA.

Reader Service No. 103

Perfectprep 96 Vac

Eppendorf www.eppendorf.com

Rapid isolation of plasmid DNA from 96 wells

Optimized for use in manual and automatic sequencing, the Perfectprep 96 Vac kit can be used for all follow-on downstream applications. The system works with membrane plates and vacuum technology, with bacterial cultures being grown in the 96-well plate. Modified alkaline lysis combined with

optimized DNA-binding groups facilitates rapid purification of plasmid. The plasmid DNA can be used immediately after elution, with no further precipitation necessary, and is free of salts and alcohol residue. Elution volume is approximately 70 µl.

Reader Service No. 104

TRANSFAC

Biobase www.biobase.de

Consort with SNPs

Data produced by the SNP Consortium is now being integrated into the TRANSFAC database of transcription factors. It is believed that the human body contains three million such gene variations on the genome, of which almost one-third have been identified. SNPs may play a role in the development of diseases such as diabetes, Alzheimer's or cancer. The TRANSFAC database currently contains nearly 24,000 entries. More about SNPs can be found at <http://snp.cshl.org/>.

Reader Service No. 105

ADVERTISEMENTS

Custom Peptides & Polyclonal Antibodies

Please contact us about complete services for your custom peptide and polyclonal antibody needs.

Research Genetics

U.S. or Canada 800-533-4363

FAX 256-536-9016

Check out our homepage at <http://www.resgen.com>

READER ENQUIRY NO. 39

DIVERSE SMALL MOLECULES READY FOR SCREENING



10K-50K
Compounds
in Plates

- Next Day Delivery*
- Competitively Priced

*Some Restrictions Apply

Chembridge Corporation • 800.964.6143
858.451.7400 • 858.451.7401 fax • E-mail: chem@chembridge.com

www.chembridge.com

READER ENQUIRY NO. 60

new on the market

ICAT Reagent Technology

Applied Biosystems

www.appliedbiosystems.com

A new approach to quantitative proteomics

ICAT Reagent Technology, developed by Ruedi Aebersold at the University of Washington, is now available for applications in proteomics. These isotope coded affinity tag reagents enable researchers to identify and quantify a broad range of proteins, including membrane and low-abundance proteins. The reagents provide a new method for preparing complex samples for mass spectrometry analysis to identify component proteins and determine their relative expression levels.

Reader Service No. 106

Apoptosis marker

Upstate Biotechnology

www.upstatebiotech.com

A universal marker of double-stranded DNA breaks and apoptosis

Double-stranded (ds) DNA breaks can be induced by radiation or chemicals, and are a characteristic event during apoptosis, and assays of DNA fragmentation are widely used to monitor apoptotic cell death. Phosphorylation of Ser 139 in the conserved C-

terminal motif of histone H2AX, a ubiquitous member of the H2A histone family, occurs within seconds of DNA fragmentation induced by various agents. Phosphorylation of H2AX occurs well before the appearance of internucleosomal DNA fragments and the externalization of phosphatidylserine to the outer membrane. Staining of cells with antibody to phospho-H2AX (Ser 139) (Upstate catalogue number 07-164) thus affords sensitive detection of dsDNA breaks.

Reader Service No. 107

20S Proteasome Assay Kit

BIOMOL Research Laboratories, Inc.

www.biomol.com

Inhibitor screening

The 20S Proteasome Assay Kit is a complete assay system for measuring the chymotrypsin-like protease activity of purified 20S proteasome. The kit is fluorogenic, non-radioactive and provides real-time kinetics of substrate cleavage; it can be used to screen compound libraries in the search for new inhibitors or to perform detailed kinetic studies on known inhibitors. Enough enzyme, substrate and inhibitors (purified human erythrocyte 20S proteasome, Suc-LLVY-AMC fluorogenic peptide substrate and lactacystin, respectively) to run 96 assays

are supplied with each kit. The 20S proteasome, the catalytic core of the larger 26S proteasome, degrades ubiquitinated proteins and is involved in many diverse cellular functions including cell-cycle progression, antigen presentation and activation of transcription factors. As such, it is a newly recognized molecular target for the discovery of novel therapeutics for diseases such as cancer and inflammatory diseases.

Reader Service No. 108

Lyotrap-Ultra

LTE Scientific

www.lte-scientific.co.uk

A large-capacity freeze-dryer

With the introduction of this new freezer, the range of Lyotrap freeze-dryers now consists of four models with ice capacities of from 3 to 18 kg. The 18-kg ice capacity, the Lyotrap-Ultra, is the largest freeze-dryer in the range, suitable for laboratory and pilot processes where high product volumes and long running times between defrosts are required. The floor-standing device is microprocessor controlled, and incorporates an electric defrost facility, with LCD temperature and vacuum parameter display.

Reader Service No. 109

These notes are compiled in the Nature office from information provided by the manufacturers.

ADVERTISEMENT